Risk Analysis for Bank

Mingdong HE

Introduction

In this analysis, we investigated the relationship between the credit score and various other measurements. We use *Train.txt* to build a model and then to predict credit scores based on it. We also classify the individuals from *Test.txt* into two groups and we get the proportion of the individuals which are correctly classified by our model. Prediction intervals are also given.

Exploratory Analysis

This data set is composed of 21 measured data about loan applicants. For each applicant, the credit score has been measured, along with the following 20 variables:

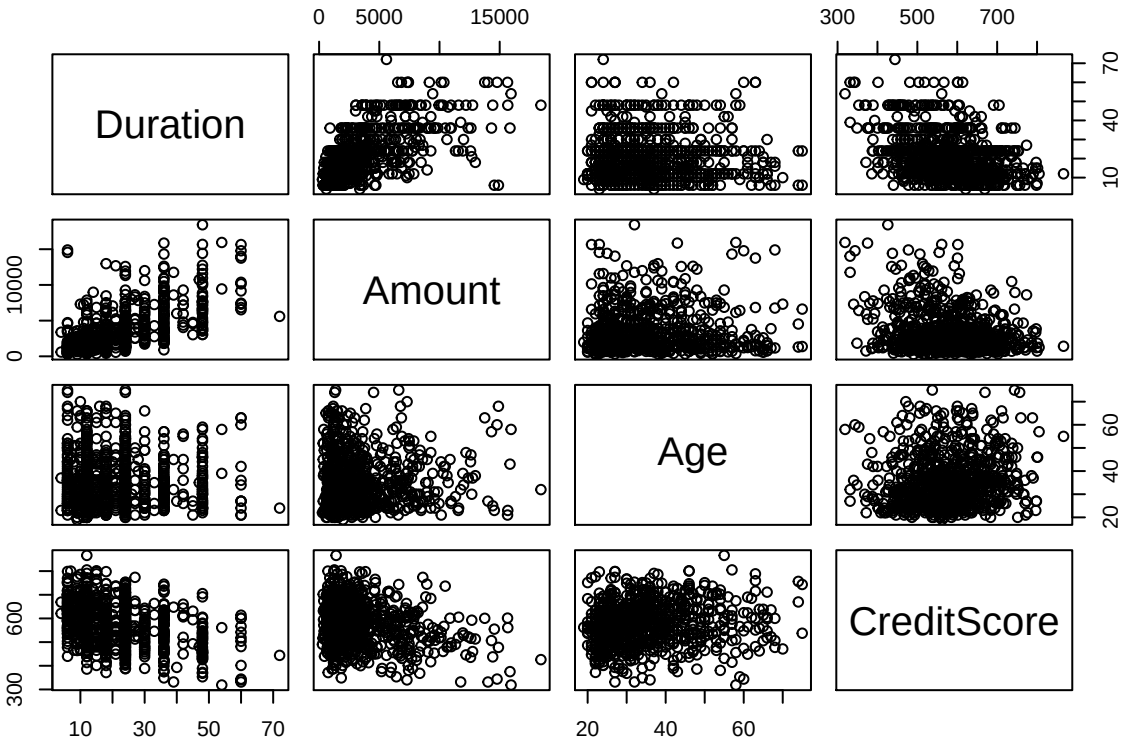
Status, Duration, History, Purpose, Amount, Savings, Employment, Disposable, Personal, OtherParties, Residence, Property, Age, Plans, Housing, Existing, Job, Dependants, Telephone, Foreign, CreditScore.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## 'data.frame':    800 obs. of  21 variables:
##  $ Status      : Factor w/ 4 levels "Large","Negative",..: 2 4 3 2 3 3 4 4 4 2 ...
##  $ Duration    : int  6 48 12 42 36 24 36 30 12 48 ...
##  $ History     : Factor w/ 5 levels "A","B","C","D",..: 5 3 5 3 3 3 3 5 3 3 ...
##  $ Purpose     : Factor w/ 10 levels "Business","Domestic",..: 8 8 3 4 3 4 10 5 5 1 ...
##  $ Amount      : int  1169 5951 2096 7882 9055 2835 6948 5234 1295 4308 ...
##  $ Savings     : Factor w/ 5 levels "Large","Low",..: 4 2 2 2 4 1 2 2 2 2 ...
##  $ Employment  : Factor w/ 5 levels "Long","Medium",..: 5 2 1 1 2 5 2 4 3 3 ...
##  $ Disposable  : Factor w/ 4 levels "1","2","3","4": 4 2 2 2 2 3 2 4 3 3 ...
##  $ Personal    : Factor w/ 4 levels "F:DivSepMar",..: 4 1 4 4 4 4 4 2 1 1 ...
##  $ OtherParties: Factor w/ 3 levels "Coapp","Guarantor",..: 3 3 3 2 3 3 3 3 3 3 ...
##  $ Residence   : Factor w/ 4 levels "1","2","3","4": 4 2 3 4 4 4 2 2 1 4 ...
##  $ Property    : Factor w/ 4 levels "Car","House",..: 2 2 2 4 3 4 1 1 1 4 ...
```

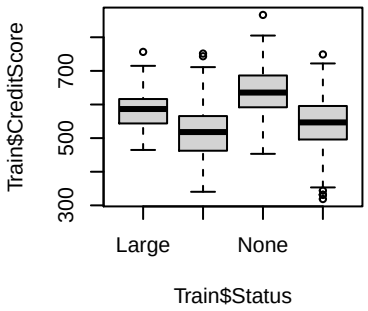
```
##  $ Age         : int  67 22 49 45 35 53 35 28 25 24 ...
##  $ Plans       : Factor w/ 3 levels "Bank","None",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Housing     : Factor w/ 3 levels "Own","Rent","RentFree": 1 1 1 3 3 1 2 1 2 2 ...
##  $ Existing    : Factor w/ 4 levels "1","2","3","4": 2 1 1 1 1 1 1 2 1 1 ...
##  $ Job         : Factor w/ 4 levels "Management:Self",..: 2 2 4 2 4 2 1 1 2 2 ...
##  $ Dependants  : Factor w/ 2 levels "1","2": 1 1 2 2 2 1 1 1 1 1 ...
##  $ Telephone   : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 1 2 1 1 1 ...
##  $ Foreign     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ CreditScore : num  636 372 678 613 575 ...
```

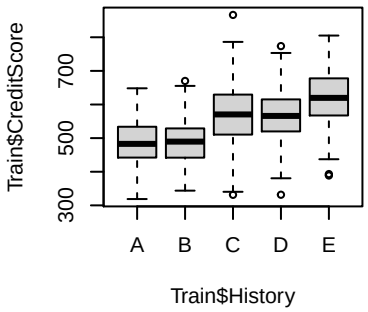
We will only treat Duration, Amount, Age as numerical variables while others will be treated as factors, since other types of variables are either factors or with a very small number of distinct possibilities they can take.

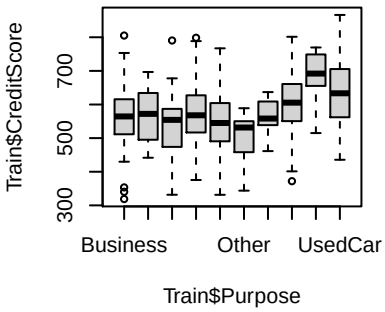
Let's plot the pairwise scatter plots for continuous measurements.

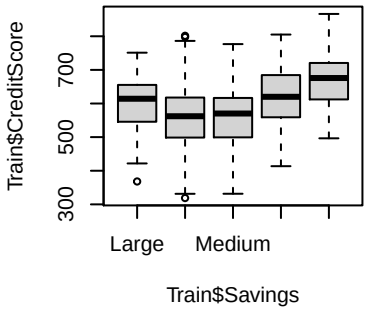


Let's use boxplots for factors:

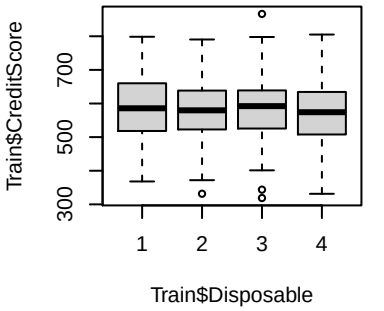

Train$CreditScore
300
500
700
Large
None
Train$Status

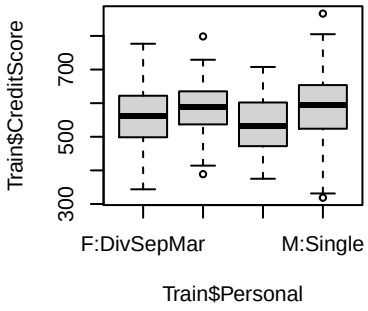

Train$CreditScore
300
500
700
A
B
C
D
E
Train$History

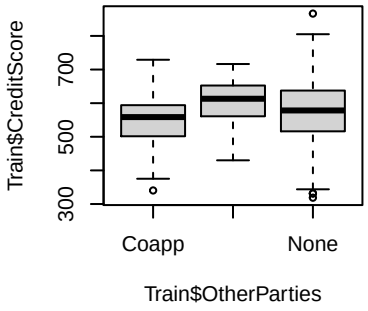

Train$CreditScore
300
500
700
Business
Other
UsedCar
Train$Purpose

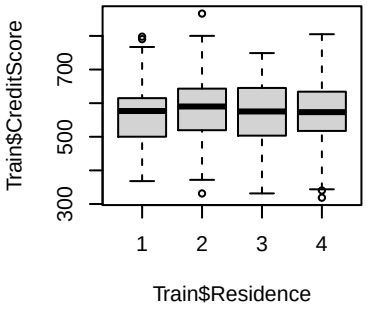

Train$CreditScore
300
500
700
Large
Medium
Train$Savings

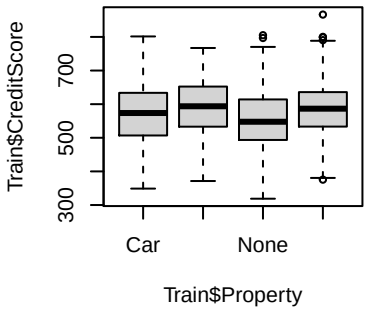

Train$CreditScore
300
500
700
Long
Short
VeryLong
Train$Employment

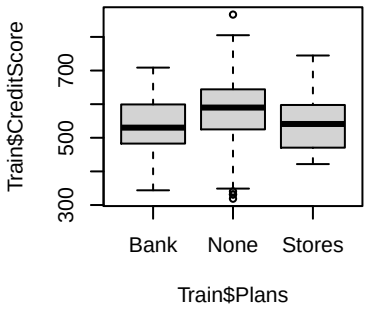

Train$CreditScore
300
500
700
1
2
3
4
Train$Disposable


Train$CreditScore
300
500
700
F:DivSepMar
M:Single
Train$Personal

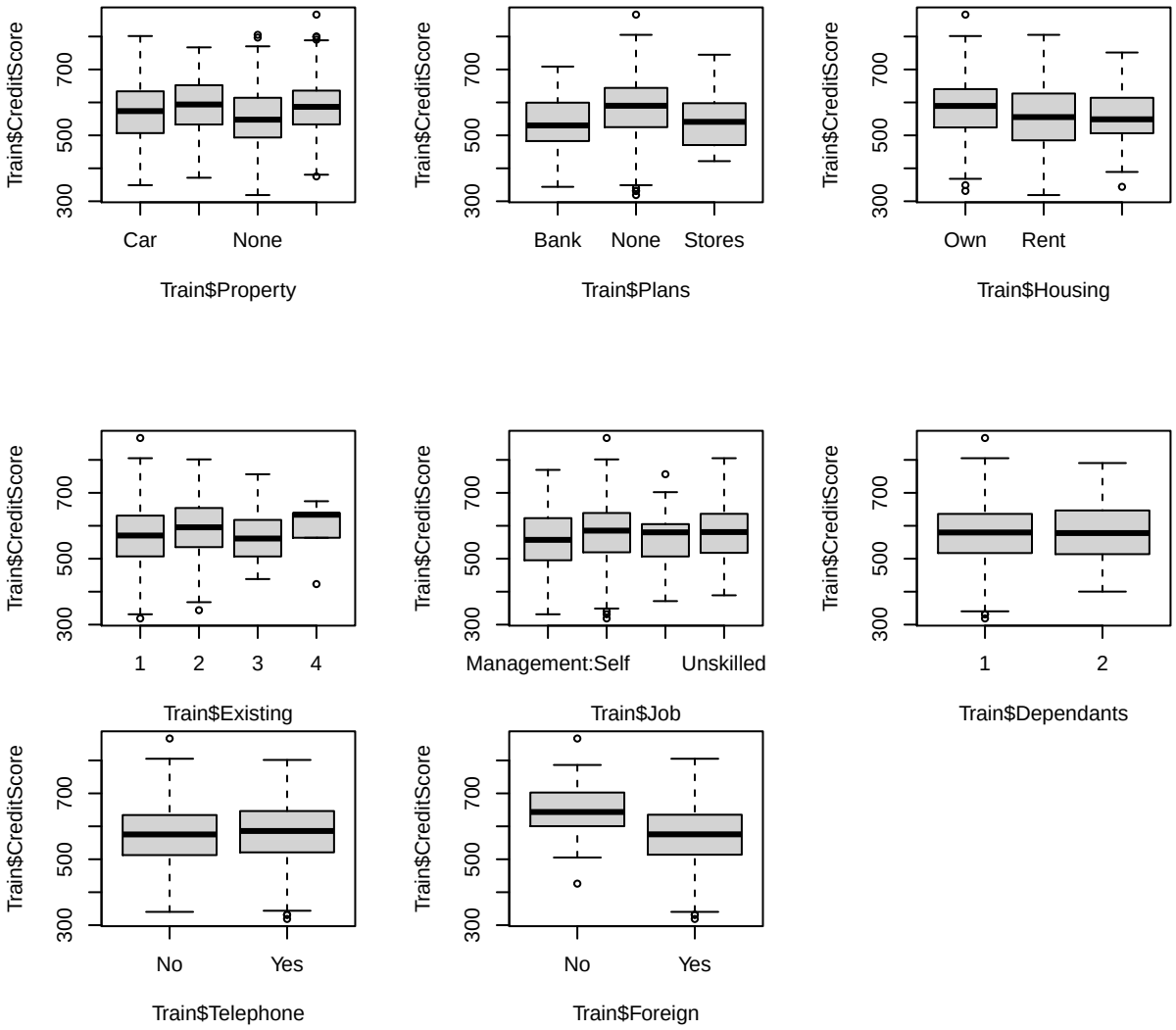

Train$CreditScore
300
500
700
Coapp
None
Train$OtherParties


Train$CreditScore
300
500
700
1
2
3
4
Train$Residence


Train$CreditScore
300
500
700
Car
None
Train$Property


Train$CreditScore
300
500
700
Bank
None
Stores
Train$Plans


Train$CreditScore
300
500
700
Own
Rent
Train$Housing

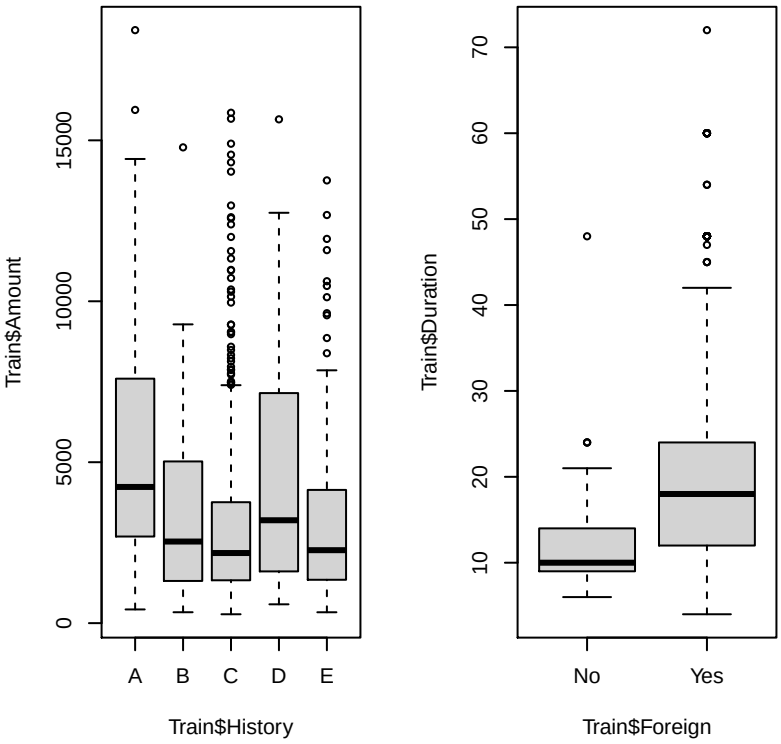


Interpretation:

The pariwise scatter plot shows some relationships between each two variables. Focusing on the last row of figures, the associations between CreditScore and all the other variables can be seen. Applicant with higher Duration and Amount seems to have lower CreditScore while applicant with higher Age seems to have higher CreditScore. However, it also is important to note that there exists collinearity between the potential explanatory variables. For instance, applicant with higher Duration tends to have higher Amount, which shows that there seems to be a strong positive association between Duration and Amount. Therefore, it could be the fact that only one or two of these variables are required to explain differences in CreditScore. We can also see that data tend to cluster, especially for Amount and Age and we see a couple of extreme Amount points, which distorts the graph and make the general trend between Amount and other variables harder to determine from the plots. Taking *log(Amount)* might drastically reduces this effect.

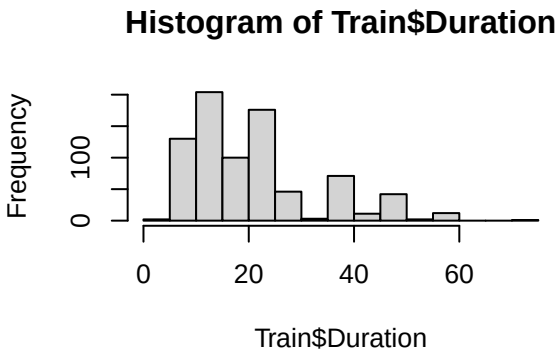
From the boxplots, we can see that applicant with Status (None) or History (E) or Purpose (UsedCar) or Savings (VeryLarge), Employment (VeryLong) or Disposable (3) or Personal (Single) or OtherParties (Guarantor) or Residence (2) or Property (House), Plans (None) or Housing (Own) or Existing (2) or Job (Skilled) or Dependants (1) or Telephone (Yes) or Foreign (No) tends to have higher a CreditScore. However, it seems that some factors like Employment, Disposable, Residence, Existing, Job, Dependants and Telephone do not have a significant affect on the CreditScore so the differences are very small. Moreover, there still exists correlations between these factors and continuous variables. For instance, applicant with worse previous history seems to have higher credit score, which could be explained by the correlation between History and

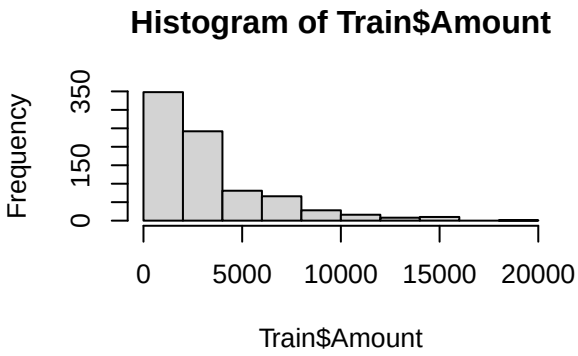
Amount, i.e. applicant with bad previous loan history can not get much loan since banks are reluctant to lend money to him. We also see that foreign applicant has longer duration of requested loan.

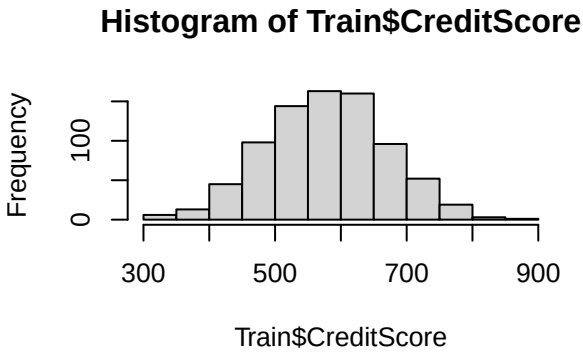


Notice that this is just based on our eyeball so it is not persuasive. We need to use quantitative methods to find an appropriate model.

Let's also check the distribution for continuous variables.







Interpretation:

From the histogram, besides the good distribution of CreditScore, the Duration, Amount and Age are all positively skewed. In order to rectify the skewness, we apply log transformation for them by Mosteller and Tukey's bulging rule.

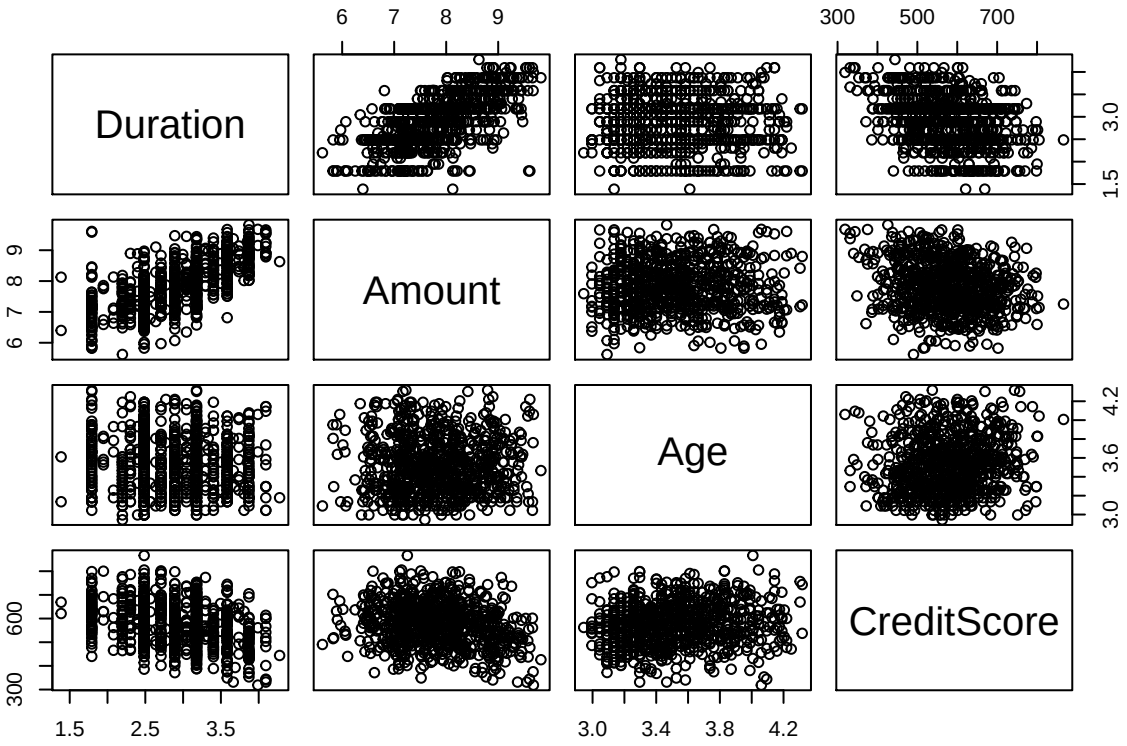
$Duration \rightarrow log(Duration)$

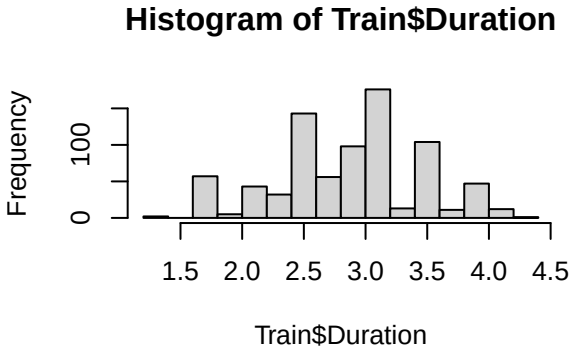
$Amount \rightarrow log(Amount)$

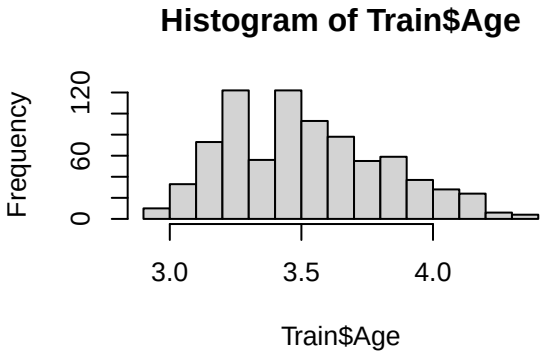
$Age \rightarrow log(Age)$

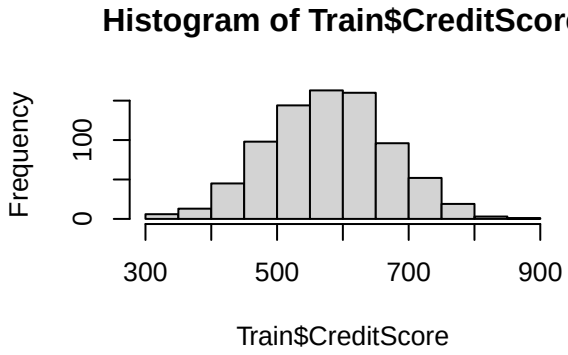
(Please note that the data we use later has already been transformed as above, which will not be shown in the summary output).

Then we plot to have a look again for our transformed data.

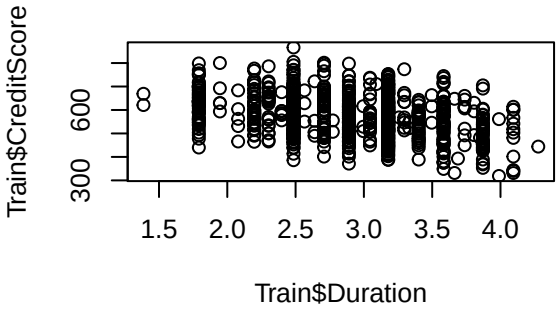


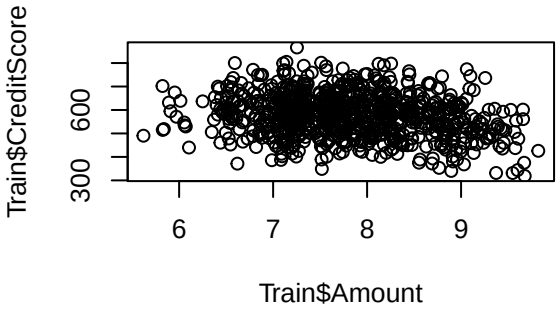


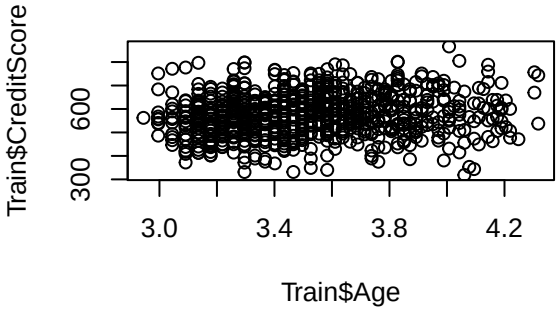




We could see that the positive skewness has been improved by transformation.







Interpretation:

From the plot, we can see the general associations:

- negative association between $log(Duration)$ and $CreditScore$
- negative association between $log(Amount)$ and $CreditScore$
- positive association between $log(Age)$ and $CreditScore$

We also see that there might exist unusual points which might provide us more information after we remove them.

Model fitting

We start by fitting the model with all possible explanatory variables included, called *full model*.

```
##
## Call:
## lm(formula = CreditScore ~ ., data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.578  -31.468   -2.637   32.134  156.872
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           778.1479    42.5276  18.297  < 2e-16 ***
## StatusNegative        -50.4976     8.2933  -6.089 1.82e-09 ***
## StatusNone             38.5923     8.0445   4.797 1.94e-06 ***
## StatusSmall           -30.6308     8.3314  -3.677 0.000253 ***
## Duration              -29.7928     4.6792  -6.367 3.37e-10 ***
## HistoryB                2.3532    12.7499   0.185 0.853618
## HistoryC               50.3760     9.9269   5.075 4.91e-07 ***
```

```
## HistoryD               50.5315    10.9453   4.617 4.59e-06 ***
## HistoryE               84.1745     9.9568   8.454  < 2e-16 ***
## PurposeDomestic       -23.4216    17.3277  -1.352 0.176887
## PurposeEducation      -55.8165    10.3391  -5.399 9.04e-08 ***
## PurposeFurniture       -1.9622     7.5213  -0.261 0.794256
## PurposeNewCar         -46.8229     7.1297  -6.567 9.61e-11 ***
## PurposeOther          -18.9078    17.1478  -1.103 0.270541
## PurposeRepairs        -24.1195    13.6494  -1.767 0.077625 .
## PurposeTelevision       1.3915     7.0057   0.199 0.842610
## PurposeTraining        54.9928    22.1630   2.481 0.013310 *
## PurposeUsedCar         36.1735     8.3833   4.315 1.81e-05 ***
## Amount                -23.1024     3.9035  -5.918 4.96e-09 ***
## SavingsLow            -11.9734     7.9806  -1.500 0.133955
## SavingsMedium           0.2690     9.4066   0.029 0.977196
## SavingsUnknown         26.5590     8.7046   3.051 0.002361 **
## SavingsVeryLarge       64.8429    11.5184   5.629 2.56e-08 ***
## EmploymentMedium        3.0379     5.6086   0.542 0.588226
## EmploymentShort         2.9163     6.3078   0.462 0.643978
## EmploymentUnemployed   -5.7077     9.6405  -0.592 0.553994
## EmploymentVeryLong      7.2925     6.2017   1.176 0.240012
## Disposable2           -12.6185     6.6312  -1.903 0.057437 .
## Disposable3           -26.6862     7.0979  -3.760 0.000183 ***
## Disposable4           -42.4973     6.8905  -6.167 1.14e-09 ***
## PersonalF:Single        9.0694     6.9040   1.314 0.189372
## PersonalM:DivSepMar   -13.8073     8.8004  -1.569 0.117087
## PersonalM:Single       30.5786     4.5496   6.721 3.59e-11 ***
## OtherPartiesGuarantor  62.8929    12.4525   5.051 5.54e-07 ***
## OtherPartiesNone       10.7948     9.0777   1.189 0.234759
## Residence2            -11.5905     6.1735  -1.877 0.060844 .
## Residence3             -8.6865     7.0604  -1.230 0.218967
## Residence4             -9.0506     6.2039  -1.459 0.145029
## PropertyHouse         -10.6165     5.0724  -2.093 0.036688 *
## PropertyNone           -7.0200     8.4944  -0.826 0.408830
## PropertySavings         0.9823     4.9688   0.198 0.843334
## Age                    15.8112     7.4336   2.127 0.033750 *
## PlansNone              33.1283     5.2681   6.288 5.46e-10 ***
## PlansStores            13.2831     9.7897   1.357 0.175242
## HousingRent           -15.6789     5.1547  -3.042 0.002435 **
## HousingRentFree         2.4301     9.6885   0.251 0.802016
## Existing2               0.6969     4.9211   0.142 0.887422
## Existing3             -11.6177    12.2024  -0.952 0.341363
## Existing4              -8.7326    22.8863  -0.382 0.702895
## JobSkilled              1.9408     6.0955   0.318 0.750265
## JobUnemployed           6.1215    14.7725   0.414 0.678709
## JobUnskilled            6.4099     7.5020   0.854 0.393142
## Dependants2            -2.3503     5.3310  -0.441 0.659428
## TelephoneYes           13.6043     4.1563   3.273 0.001113 **
## ForeignYes            -55.3346     9.7670  -5.665 2.10e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.49 on 745 degrees of freedom
## Multiple R-squared:  0.7233, Adjusted R-squared:  0.7033
## F-statistic: 36.07 on 54 and 745 DF,  p-value: < 2.2e-16
```

```
## [1] "fit_full$adj.r.squared = 0.703252"
```

Interpretation:

We see from the summary output that the $R^2_{adj} = 0.703252$ suggests a relatively good fit.

Such small $p - values$ for Status tell us to reject H_0, which is an evidence that it has strong association with CreditScore. This could also been seen apparently from previous exploratory plots. As well as for History, applicant with History (C) or Hisotory (D) or History (E) tends to have a much higher credit score than History (A).

For continuous variables, Duration and Amount are strongly associated with CreditScore by their $p-values < 0.001$ while Age also has a very small $p - value = 0.03375 < 0.05$, so we reject H_0 at 5% level. From the exploratory plots, applicant with longer Duration or larger Amount has lower CreditScore, while applicant with higher Age has a higher CreditScore.

However, variables such as Employment, Residence, Existing, Job and Dependants have $p - values$ greater than 0.1, which are not significant at the 10% level after accounting for other factors. So there is no evidence that these variables are associated with CreditScore. The $p - values$ for Residence2 and PurposeRepairs are greater than 0.05 but less than 0.1, which could be considered as mildly significant. From the exploratory plots, we see that applicant with 2 years in current residence have a higher credit score than others in full years residence while applicant with purpose of Repairs have a higher credit score than other purposes.

The full model is not an ideal model since least significant variables are also included and it is too complex, which might cause over-fit and affect the accuracy for prediction. So the next step is to see which variables should be retained in order to find the most important predictors.

Model selection

Due to a large number of variables, best subsets regression is not a good choice since there will be 2^{20} models to be fitted. Therefore, we try using the stepwise regression to optimize our variables selection in our model by using the AIC as the criterion.

Stepwise regression

Let's fit the full model and use the backward stepwise regression.

```
## Start:  AIC=6295.92
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Employment + Disposable + Personal + OtherParties +
##     Residence + Property + Age + Plans + Housing + Existing +
##     Job + Dependants + Telephone + Foreign
##
##                Df Sum of Sq     RSS    AIC
## - Employment    4      6721 1831580 6290.9
## - Job           3      2609 1827468 6291.1
## - Existing      3      2921 1827780 6291.2
## - Residence     3      8754 1833613 6293.7
## - Dependants    1       476 1825335 6294.1
## <none>                      1824859 6295.9
## - Property      3     15268 1840128 6296.6
## - Age           1     11082 1835941 6298.8
## - Housing       2     24818 1849677 6302.7
## - Telephone     1     26243 1851102 6305.3
## - Foreign       1     78622 1903481 6327.7
## - OtherParties  2     85881 1910741 6328.7
## - Amount        1     85797 1910656 6330.7
```

```
## - Plans         2    101747 1926606 6335.3
## - Duration      1     99299 1924158 6336.3
## - Disposable    3    126664 1951523 6343.6
## - Personal      3    145829 1970689 6351.4
## - Savings       4    281432 2106291 6402.7
## - History       4    296658 2121518 6408.4
## - Purpose       9    526935 2351794 6480.9
## - Status        3    988780 2813639 6636.3
##
## Step:  AIC=6290.86
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Residence +
##     Property + Age + Plans + Housing + Existing + Job + Dependants +
##     Telephone + Foreign
##
##                Df Sum of Sq     RSS    AIC
## - Existing      3      2676 1834256 6286.0
## - Job           3      3327 1834907 6286.3
## - Residence     3      9103 1840683 6288.8
## - Dependants    1       351 1831931 6289.0
## <none>                      1831580 6290.9
## - Property      3     14669 1846249 6291.2
## - Age           1     15893 1847473 6295.8
## - Housing       2     24291 1855871 6297.4
## - Telephone     1     26006 1857586 6300.1
## - Foreign       1     78974 1910555 6322.6
## - OtherParties  2     85705 1917285 6323.4
## - Amount        1     88030 1919610 6326.4
## - Plans         2    100146 1931726 6329.5
## - Duration      1     98149 1929729 6330.6
## - Disposable    3    125143 1956723 6337.7
## - Personal      3    152284 1983865 6348.8
## - Savings       4    290564 2122144 6400.7
## - History       4    301439 2133019 6404.7
## - Purpose       9    530002 2361582 6476.2
## - Status        3   1000070 2831650 6633.4
##
## Step:  AIC=6286.03
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Residence +
##     Property + Age + Plans + Housing + Job + Dependants + Telephone +
##     Foreign
##
##                Df Sum of Sq     RSS    AIC
## - Job           3      3915 1838172 6281.7
## - Residence     3      9102 1843359 6284.0
## - Dependants    1       422 1834678 6284.2
## <none>                      1834256 6286.0
## - Property      3     15287 1849544 6286.7
## - Age           1     15183 1849439 6290.6
## - Housing       2     24230 1858486 6292.5
## - Telephone     1     25083 1859339 6294.9
## - Foreign       1     78633 1912890 6317.6
## - OtherParties  2     85356 1919612 6318.4
```

```
## - Amount        1     89471 1923727 6322.1
## - Duration      1     97300 1931556 6325.4
## - Plans         2    104800 1939056 6326.5
## - Disposable    3    123624 1957881 6332.2
## - Personal      3    150738 1984994 6343.2
## - Savings       4    292897 2127153 6396.5
## - History       4    340631 2174887 6414.3
## - Purpose       9    529892 2364149 6471.1
## - Status        3   1008684 2842940 6630.6
##
## Step:  AIC=6281.74
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Residence +
##     Property + Age + Plans + Housing + Dependants + Telephone +
##     Foreign
##
##                Df Sum of Sq     RSS    AIC
## - Residence     3      8632 1846803 6279.5
## - Dependants    1       169 1838340 6279.8
## - Property      3     13745 1851917 6281.7
## <none>                      1838172 6281.7
## - Age           1     15465 1853636 6286.4
## - Housing       2     24777 1862948 6288.4
## - Telephone     1     22655 1860827 6289.5
## - Foreign       1     80681 1918853 6314.1
## - OtherParties  2     88444 1926616 6315.3
## - Duration      1     96615 1934787 6320.7
## - Amount        1     98442 1936613 6321.5
## - Plans         2    103740 1941912 6321.7
## - Disposable    3    132072 1970243 6331.2
## - Personal      3    151437 1989608 6339.1
## - Savings       4    295645 2133816 6393.0
## - History       4    343312 2181483 6410.7
## - Purpose       9    534110 2372281 6467.8
## - Status        3   1014374 2852546 6627.3
##
## Step:  AIC=6279.48
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Property +
##     Age + Plans + Housing + Dependants + Telephone + Foreign
##
##                Df Sum of Sq     RSS    AIC
## - Dependants    1       252 1847056 6277.6
## - Property      3     12030 1858833 6278.7
## <none>                      1846803 6279.5
## - Age           1     15658 1862461 6284.2
## - Housing       2     27489 1874293 6287.3
## - Telephone     1     23036 1869839 6287.4
## - Foreign       1     79864 1926667 6311.4
## - OtherParties  2     84991 1931794 6311.5
## - Duration      1     95945 1942748 6318.0
## - Plans         2    102742 1949545 6318.8
## - Amount        1    102240 1949043 6320.6
## - Disposable    3    133672 1980475 6329.4
```

```
## - Personal      3    148334 1995138 6335.3
## - Savings       4    291778 2138582 6388.8
## - History       4    339799 2186602 6406.6
## - Purpose       9    539302 2386106 6466.4
## - Status        3   1019646 2866449 6625.2
##
## Step:  AIC=6277.59
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Property +
##     Age + Plans + Housing + Telephone + Foreign
##
##                Df Sum of Sq     RSS    AIC
## - Property      3     12294 1859350 6276.9
## <none>                      1847056 6277.6
## - Age           1     15533 1862588 6282.3
## - Housing       2     27512 1874567 6285.4
## - Telephone     1     23210 1870266 6285.6
## - Foreign       1     79678 1926733 6309.4
## - OtherParties  2     84821 1931877 6309.5
## - Duration      1     95756 1942811 6316.0
## - Plans         2    103372 1950428 6317.2
## - Amount        1    102221 1949277 6318.7
## - Disposable    3    133442 1980497 6327.4
## - Personal      3    154229 2001285 6335.8
## - Savings       4    291634 2138690 6386.9
## - History       4    341249 2188304 6405.2
## - Purpose       9    543096 2390152 6465.8
## - Status        3   1019402 2866457 6623.2
##
## Step:  AIC=6276.9
## CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Age + Plans +
##     Housing + Telephone + Foreign
##
##                Df Sum of Sq     RSS    AIC
## <none>                      1859350 6276.9
## - Age           1     14215 1873565 6281.0
## - Housing       2     29256 1888606 6285.4
## - Telephone     1     25915 1885265 6286.0
## - OtherParties  2     81759 1941109 6307.3
## - Foreign       1     77593 1936943 6307.6
## - Duration      1     92007 1951357 6313.5
## - Plans         2    100841 1960191 6315.2
## - Amount        1    100075 1959425 6316.8
## - Disposable    3    132982 1992331 6326.2
## - Personal      3    156712 2016062 6335.6
## - Savings       4    291121 2150471 6385.3
## - History       4    342290 2201640 6404.1
## - Purpose       9    552050 2411400 6466.9
## - Status        3   1018039 2877389 6620.2
```

Interpretation:

The AIC value is 6295.92 for full model, after backward stepwise regression, the AIC has been reduced to 6276.9 and the current model is the relatively optimized model and let's check its summary output.

```
##
## Call:
## lm(formula = CreditScore ~ Status + Duration + History + Purpose +
##     Amount + Savings + Disposable + Personal + OtherParties +
##     Age + Plans + Housing + Telephone + Foreign, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.51  -31.67   -2.61   32.64  161.55
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            773.15372   39.30008  19.673  < 2e-16 ***
## StatusNegative         -50.87728    8.13959  -6.251 6.80e-10 ***
## StatusNone              37.58202    7.92566   4.742 2.53e-06 ***
## StatusSmall            -31.32648    8.24029  -3.802 0.000155 ***
## Duration               -28.12235    4.57978  -6.141 1.32e-09 ***
## HistoryB                -0.84901   12.38169  -0.069 0.945350
## HistoryC                48.63893    9.46070   5.141 3.47e-07 ***
## HistoryD                49.07111   10.79740   4.545 6.40e-06 ***
## HistoryE                82.21082    9.79985   8.389 2.37e-16 ***
## PurposeDomestic        -24.81449   17.11427  -1.450 0.147490
## PurposeEducation       -52.96452   10.15649  -5.215 2.37e-07 ***
## PurposeFurniture         0.15346    7.37124   0.021 0.983395
## PurposeNewCar          -46.23655    7.04418  -6.564 9.69e-11 ***
## PurposeOther           -19.87584   16.57907  -1.199 0.230958
## PurposeRepairs         -25.63507   13.45493  -1.905 0.057123 .
## PurposeTelevision        2.30334    6.93365   0.332 0.739832
## PurposeTraining         59.37339   21.67226   2.740 0.006295 **
## PurposeUsedCar          37.90101    8.21757   4.612 4.67e-06 ***
## Amount                 -23.88738    3.73000  -6.404 2.64e-10 ***
## SavingsLow             -13.14080    7.86423  -1.671 0.095140 .
## SavingsMedium            0.06139    9.28806   0.007 0.994728
## SavingsUnknown          25.83608    8.59536   3.006 0.002736 **
## SavingsVeryLarge        62.47214   11.29406   5.531 4.37e-08 ***
## Disposable2            -11.97283    6.48331  -1.847 0.065176 .
## Disposable3            -26.72090    6.94964  -3.845 0.000131 ***
## Disposable4            -41.89259    6.66438  -6.286 5.48e-10 ***
## PersonalF:Single         7.94241    6.75865   1.175 0.240303
## PersonalM:DivSepMar    -13.30743    8.70654  -1.528 0.126818
## PersonalM:Single        30.06153    4.33588   6.933 8.78e-12 ***
## OtherPartiesGuarantor   61.12648   12.15645   5.028 6.17e-07 ***
## OtherPartiesNone        11.94918    8.90652   1.342 0.180118
## Age                     16.27202    6.74177   2.414 0.016030 *
## PlansNone               32.27987    5.16452   6.250 6.81e-10 ***
## PlansStores             11.89419    9.68139   1.229 0.219615
## HousingRent            -16.91325    4.89081  -3.458 0.000574 ***
## HousingRentFree         -3.76808    6.15605  -0.612 0.540658
## TelephoneYes            12.64238    3.87933   3.259 0.001168 **
## ForeignYes             -54.42600    9.65159  -5.639 2.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.4 on 762 degrees of freedom
```

```
## Multiple R-squared:  0.7181, Adjusted R-squared:  0.7044
## F-statistic: 52.46 on 37 and 762 DF,  p-value: < 2.2e-16
```

Interpretation:

The result tells us that we could use the model:

$CreditScore \sim Status + log(Duration) + History + Purpose + log(Amount) + Savings + Disposable + Personal + OtherParties + log(Age) + Plans + Housing + Telephone + Foreign.$ Let's call it *fit1*.

```
## [1] "fit1$adj.r.squared = 0.704389"
```

And the corresponding coefficients of this model are:

```
##           (Intercept)        StatusNegative            StatusNone
##          773.15371693          -50.87727687           37.58202440
##           StatusSmall              Duration              HistoryB
##          -31.32648063          -28.12234742           -0.84900773
##              HistoryC              HistoryD              HistoryE
##           48.63892583           49.07110935           82.21081593
##       PurposeDomestic     PurposeEducation     PurposeFurniture
##          -24.81448928          -52.96452379            0.15346238
##         PurposeNewCar         PurposeOther        PurposeRepairs
##          -46.23655326          -19.87584346          -25.63507120
##     PurposeTelevision      PurposeTraining        PurposeUsedCar
##            2.30333989           59.37338536           37.90101249
##                Amount            SavingsLow         SavingsMedium
##          -23.88737916          -13.14079907            0.06138926
##        SavingsUnknown     SavingsVeryLarge           Disposable2
##           25.83607730           62.47214238          -11.97282733
##           Disposable3           Disposable4     PersonalF:Single
##          -26.72090375          -41.89259098            7.94241256
##   PersonalM:DivSepMar     PersonalM:Single OtherPartiesGuarantor
##          -13.30742764           30.06152846           61.12647839
##      OtherPartiesNone                   Age             PlansNone
##           11.94917690           16.27202187           32.27987009
##           PlansStores           HousingRent       HousingRentFree
##           11.89419015          -16.91325132           -3.76808462
##          TelephoneYes            ForeignYes
##           12.64238486          -54.42600184
```

We find that the least significant variables such as Employment, Residence, Existing, Job and Dependants have been removed from the full model.

We see that continuous variables $log(Duration)$, $log(Amount)$ are significant at 5% level while $log(Age)$ is significant at 10% level. Considering the exploratory plots, for Duration and Amount, there is a strong negative association between CreditScore and them while for Age, mild positive association is seen.

We also see the strong association between Status, History, Purpose, Savings, Disposable, Personal, OtherParties, Plans, Housing, Telephone, Foreign and CreditScore. Notice that this is not simultaneous interpretation, every variables is interpreted after accounting for others. Considering previous exploratory plots, we see that applicant with Status (None) or History (E) or Purpose (Training) or Savings (Verylarge) or Disposable (3) or Personal (Single), OtherParties (Guarantor) or Plans (None) or Housing (Own) or Telephone (Yes) or Foreign (Yes) has a higher CreditScore.

Note that the $R^2_{adj} = 0.704389$ still suggests a good fit, which has been increased from $R^2_{adj} = 0.703252$ of full model.

Moreover, it is worthwhile to note that the model is not guaranteed to be the best model since not every possible model is evaluated.

Let's fit the null model and use backward stepwise regression.

```
## Start:  AIC=7215.8
## CreditScore ~ 1
##
##                Df Sum of Sq     RSS    AIC
## + Status        3   2200712 4394551 6897.0
## + History       4   1022360 5572903 7089.1
## + Duration      1    753652 5841611 7120.7
## + Savings       4    727667 5867596 7130.3
## + Purpose       9    778272 5816991 7133.3
## + Amount        1    296237 6299026 7181.0
## + Plans         2    279726 6315537 7185.1
## + Personal      3    214591 6380672 7195.3
## + Foreign       1    132694 6462569 7201.5
## + Housing       2    147124 6448139 7201.8
## + Employment    4    177175 6418088 7202.0
## + Age           1    114494 6480769 7203.8
## + Property      3    118914 6476349 7207.2
## + Existing      3    114313 6480950 7207.8
## + OtherParties  2     50477 6544786 7213.7
## <none>                     6595263 7215.8
## + Job           3     47843 6547420 7216.0
## + Telephone     1     13145 6582118 7216.2
## + Disposable    3     40413 6554850 7216.9
## + Dependants    1      6029 6589234 7217.1
## + Residence     3     21928 6573335 7219.1
##
## Step:  AIC=6897.01
## CreditScore ~ Status
##
##                Df Sum of Sq     RSS    AIC
## + Duration      1    626765 3767786 6775.9
## + History       4    557015 3837536 6796.6
## + Purpose       9    433981 3960570 6831.8
## + Savings       4    353882 4040669 6837.8
## + Amount        1    259565 4134986 6850.3
## + Plans         2    198835 4195716 6864.0
## + Foreign       1    146638 4247914 6871.9
## + Personal      3    138797 4255754 6877.3
## + OtherParties  2     88580 4305972 6884.7
## + Property      3     98303 4296249 6884.9
## + Age           1     72840 4321711 6885.6
## + Housing       2     56558 4337993 6890.6
## + Employment    4     64969 4329583 6893.1
## + Disposable    3     48487 4346064 6894.1
## + Existing      3     41800 4352752 6895.4
## + Job           3     35434 4359118 6896.5
## <none>                     4394551 6897.0
## + Dependants    1      4168 4390384 6898.3
## + Telephone     1      1304 4393247 6898.8
## + Residence     3     11233 4383319 6901.0
```

```
##
## Step:  AIC=6775.91
## CreditScore ~ Status + Duration
##
##                Df Sum of Sq     RSS    AIC
## + Purpose       9    537367 3230419 6670.8
## + History       4    438103 3329683 6685.0
## + Savings       4    374012 3393774 6700.3
## + Personal      3    187698 3580088 6741.0
## + Plans         2    153705 3614081 6746.6
## + Employment    4     96931 3670856 6763.1
## + Foreign       1     67601 3700185 6763.4
## + Age           1     61188 3706599 6764.8
## + Housing       2     62149 3705638 6766.6
## + OtherParties  2     59418 3708368 6767.2
## + Existing      3     54167 3713619 6770.3
## + Telephone     1     30809 3736977 6771.3
## <none>                      3767786 6775.9
## + Property      3     24770 3743016 6776.6
## + Disposable    3     23557 3744230 6776.9
## + Dependants    1      1534 3766252 6777.6
## + Amount        1        88 3767698 6777.9
## + Job           3     10639 3757148 6779.6
## + Residence     3      4858 3762928 6780.9
##
## Step:  AIC=6670.81
## CreditScore ~ Status + Duration + Purpose
##
##                Df Sum of Sq     RSS    AIC
## + History       4    458753 2771665 6556.3
## + Savings       4    312434 2917985 6597.4
## + Personal      3    181194 3049225 6630.6
## + Plans         2    151796 3078623 6636.3
## + Foreign       1    100960 3129459 6647.4
## + Age           1     82647 3147772 6652.1
## + Housing       2     80402 3150017 6654.6
## + Employment    4     79048 3151371 6659.0
## + Existing      3     58231 3172188 6662.3
## + OtherParties  2     43629 3186790 6663.9
## + Telephone     1     24305 3206114 6666.8
## + Property      3     33824 3196595 6668.4
## <none>                      3230419 6670.8
## + Amount        1      7711 3222707 6670.9
## + Dependants    1      6524 3223895 6671.2
## + Disposable    3     19150 3211269 6672.1
## + Job           3     14028 3216390 6673.3
## + Residence     3      7506 3222913 6674.9
##
## Step:  AIC=6556.28
## CreditScore ~ Status + Duration + Purpose + History
##
##                Df Sum of Sq     RSS    AIC
## + Savings       4    282607 2489058 6478.2
## + Personal      3    155732 2615933 6516.0
```

```
## + Foreign       1    107079 2664586 6526.8
## + Plans         2     78211 2693455 6537.4
## + OtherParties  2     72063 2699603 6539.2
## + Age           1     45092 2726574 6545.2
## + Housing       2     51515 2720151 6545.3
## + Employment    4     43819 2727846 6551.5
## + Disposable    3     28885 2742781 6553.9
## + Telephone     1     14159 2757507 6554.2
## + Dependants    1     11738 2759927 6554.9
## + Amount        1      9588 2762077 6555.5
## + Property      3     21656 2750010 6556.0
## <none>                      2771665 6556.3
## + Job           3     15396 2756270 6557.8
## + Residence     3      3920 2767746 6561.1
## + Existing      3      1084 2770582 6562.0
##
## Step:  AIC=6478.24
## CreditScore ~ Status + Duration + Purpose + History + Savings
##
##                Df Sum of Sq     RSS    AIC
## + Personal      3    147554 2341505 6435.4
## + Foreign       1    101883 2387175 6446.8
## + OtherParties  2     86610 2402448 6453.9
## + Plans         2     77615 2411444 6456.9
## + Housing       2     53644 2435414 6464.8
## + Age           1     31316 2457742 6470.1
## + Disposable    3     41048 2448010 6470.9
## + Amount        1     10806 2478252 6476.8
## + Telephone     1     10730 2478328 6476.8
## + Dependants    1      9886 2479172 6477.1
## <none>                      2489058 6478.2
## + Employment    4     23382 2465677 6478.7
## + Property      3     12533 2476525 6480.2
## + Job           3     11837 2477221 6480.4
## + Residence     3      3051 2486007 6483.3
## + Existing      3       512 2488546 6484.1
##
## Step:  AIC=6435.35
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal
##
##                Df Sum of Sq     RSS    AIC
## + Foreign       1     89856 2251648 6406.0
## + Plans         2     89615 2251890 6408.1
## + OtherParties  2     85623 2255882 6409.6
## + Disposable    3     62512 2278993 6419.7
## + Housing       2     28484 2313021 6429.6
## + Amount        1     21520 2319984 6430.0
## + Age           1     11992 2329513 6433.2
## + Telephone     1      7566 2333939 6434.8
## <none>                      2341505 6435.4
## + Property      3     16404 2325101 6435.7
## + Job           3     15183 2326322 6436.1
## + Dependants    1        16 2341488 6437.3
```

```
## + Residence     3      5590 2335915 6439.4
## + Existing      3      2944 2338561 6440.3
## + Employment    4      8072 2333432 6440.6
##
## Step:  AIC=6406.05
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign
##
##                Df Sum of Sq     RSS    AIC
## + Plans         2     85249 2166399 6379.2
## + OtherParties  2     81153 2170496 6380.7
## + Disposable    3     52299 2199350 6393.2
## + Amount        1     25026 2226622 6399.1
## + Housing       2     28869 2222780 6399.7
## + Age           1     14374 2237275 6402.9
## + Telephone     1     12314 2239334 6403.7
## <none>                      2251648 6406.0
## + Property      3     13215 2238434 6407.3
## + Dependants    1        39 2251609 6408.0
## + Job           3     10771 2240878 6408.2
## + Residence     3      6892 2244756 6409.6
## + Existing      3      2461 2249187 6411.2
## + Employment    4      6672 2244976 6411.7
##
## Step:  AIC=6379.17
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans
##
##                Df Sum of Sq     RSS    AIC
## + OtherParties  2     88026 2078374 6350.0
## + Disposable    3     50543 2115856 6366.3
## + Housing       2     35178 2131221 6370.1
## + Amount        1     26064 2140335 6371.5
## + Age           1     18247 2148152 6374.4
## + Telephone     1     12304 2154095 6376.6
## <none>                      2166399 6379.2
## + Property      3     13094 2153306 6380.3
## + Dependants    1        29 2166371 6381.2
## + Job           3      9807 2156592 6381.5
## + Residence     3      7629 2158770 6382.4
## + Employment    4      8988 2157411 6383.8
## + Existing      3       465 2165935 6385.0
##
## Step:  AIC=6349.99
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans + OtherParties
##
##               Df Sum of Sq     RSS    AIC
## + Disposable  3     50008 2028366 6336.5
## + Housing     2     35748 2042626 6340.1
## + Amount      1     21816 2056558 6343.5
## + Age         1     18416 2059958 6344.9
## + Telephone   1     14562 2063811 6346.4
## <none>                    2078374 6350.0
```

```
## + Property    3     12391 2065983 6351.2
## + Residence   3     12001 2066373 6351.4
## + Dependants  1        37 2078337 6352.0
## + Job         3      4996 2073378 6354.1
## + Employment  4      8935 2069438 6354.5
## + Existing    3       784 2077590 6355.7
##
## Step:  AIC=6336.5
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans + OtherParties + Disposable
##
##              Df Sum of Sq     RSS    AIC
## + Amount      1     86779 1941587 6303.5
## + Housing     2     40388 1987978 6324.4
## + Age         1     24730 2003636 6328.7
## + Telephone   1     13096 2015270 6333.3
## <none>                    2028366 6336.5
## + Residence   3     12264 2016102 6337.7
## + Property    3     11206 2017160 6338.1
## + Dependants  1       343 2028022 6338.4
## + Employment  4     15009 2013356 6338.6
## + Job         3      3274 2025091 6341.2
## + Existing    3      2756 2025610 6341.4
##
## Step:  AIC=6303.52
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans + OtherParties + Disposable +
##     Amount
##
##              Df Sum of Sq     RSS    AIC
## + Telephone   1     32709 1908877 6291.9
## + Housing     2     35743 1905844 6292.7
## + Age         1     28116 1913470 6293.9
## <none>                    1941587 6303.5
## + Property    3     14281 1927306 6303.6
## + Dependants  1       492 1941094 6305.3
## + Residence   3      8724 1932863 6305.9
## + Employment  4     12631 1928956 6306.3
## + Job         3      1865 1939722 6308.8
## + Existing    3      1416 1940170 6308.9
##
## Step:  AIC=6291.93
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans + OtherParties + Disposable +
##     Amount + Telephone
##
##              Df Sum of Sq     RSS    AIC
## + Housing     2     35313 1873565 6281.0
## + Age         1     20271 1888606 6285.4
## <none>                    1908877 6291.9
## + Property    3     11502 1897375 6293.1
## + Dependants  1       291 1908587 6293.8
## + Residence   3      8446 1900431 6294.4
## + Employment  4     11104 1897774 6295.3
```

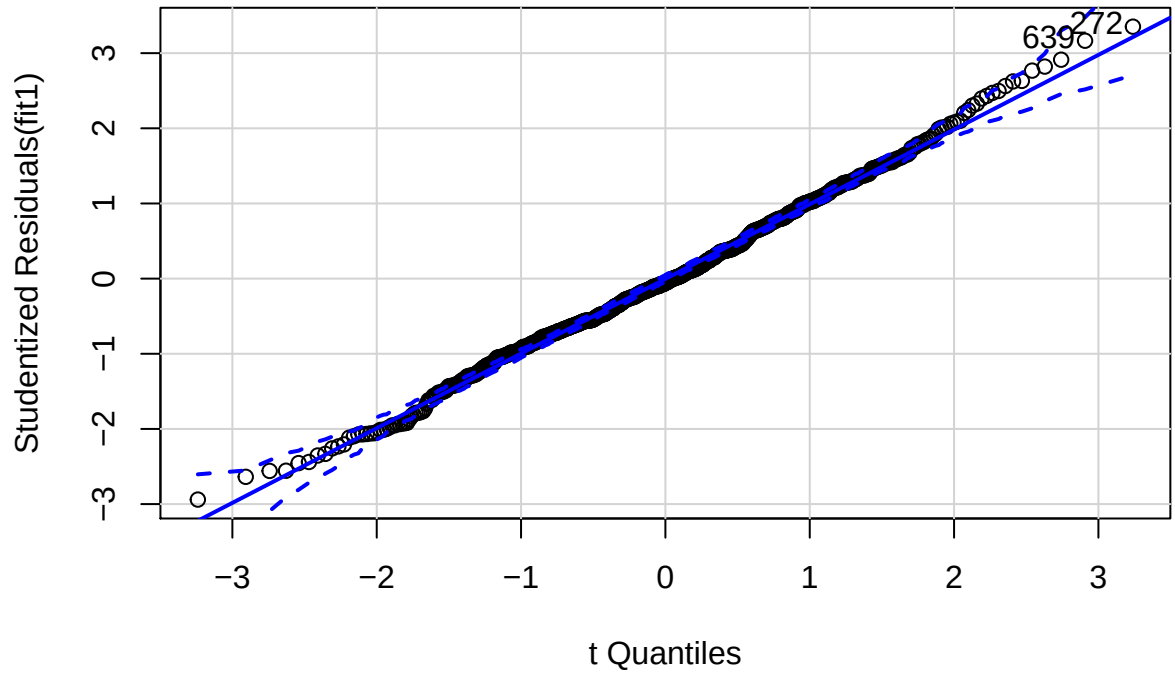
```
## + Existing    3      2702 1906175 6296.8
## + Job         3      2121 1906756 6297.0
##
## Step:  AIC=6280.99
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans + OtherParties + Disposable +
##     Amount + Telephone + Housing
##
##              Df Sum of Sq     RSS    AIC
## + Age         1   14214.8 1859350 6276.9
## <none>                    1873565 6281.0
## + Property    3   10976.3 1862588 6282.3
## + Dependants  1     314.6 1873250 6282.9
## + Residence   3    7540.8 1866024 6283.8
## + Employment  4   11442.0 1862123 6284.1
## + Existing    3    2572.5 1870992 6285.9
## + Job         3    2121.5 1871443 6286.1
##
## Step:  AIC=6276.9
## CreditScore ~ Status + Duration + Purpose + History + Savings +
##     Personal + Foreign + Plans + OtherParties + Disposable +
##     Amount + Telephone + Housing + Age
##
##              Df Sum of Sq     RSS    AIC
## <none>                    1859350 6276.9
## + Property    3   12294.3 1847056 6277.6
## + Dependants  1     516.7 1858833 6278.7
## + Residence   3    7016.2 1852334 6279.9
## + Existing    3    3792.5 1855557 6281.3
## + Employment  4    7638.2 1851712 6281.6
## + Job         3    1938.2 1857412 6282.1
##
## Call:
## lm(formula = CreditScore ~ Status + Duration + Purpose + History +
##     Savings + Personal + Foreign + Plans + OtherParties + Disposable +
##     Amount + Telephone + Housing + Age, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.51  -31.67   -2.61   32.64  161.55
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          773.15372   39.30008  19.673  < 2e-16 ***
## StatusNegative       -50.87728    8.13959  -6.251 6.80e-10 ***
## StatusNone            37.58202    7.92566   4.742 2.53e-06 ***
## StatusSmall          -31.32648    8.24029  -3.802 0.000155 ***
## Duration             -28.12235    4.57978  -6.141 1.32e-09 ***
## PurposeDomestic      -24.81449   17.11427  -1.450 0.147490
## PurposeEducation     -52.96452   10.15649  -5.215 2.37e-07 ***
## PurposeFurniture       0.15346    7.37124   0.021 0.983395
## PurposeNewCar        -46.23655    7.04418  -6.564 9.69e-11 ***
## PurposeOther         -19.87584   16.57907  -1.199 0.230958
```

```
## PurposeRepairs        -25.63507   13.45493  -1.905 0.057123 .
## PurposeTelevision       2.30334    6.93365   0.332 0.739832
## PurposeTraining        59.37339   21.67226   2.740 0.006295 **
## PurposeUsedCar         37.90101    8.21757   4.612 4.67e-06 ***
## HistoryB               -0.84901   12.38169  -0.069 0.945350
## HistoryC               48.63893    9.46070   5.141 3.47e-07 ***
## HistoryD               49.07111   10.79740   4.545 6.40e-06 ***
## HistoryE               82.21082    9.79985   8.389 2.37e-16 ***
## SavingsLow            -13.14080    7.86423  -1.671 0.095140 .
## SavingsMedium           0.06139    9.28806   0.007 0.994728
## SavingsUnknown         25.83608    8.59536   3.006 0.002736 **
## SavingsVeryLarge       62.47214   11.29406   5.531 4.37e-08 ***
## PersonalF:Single        7.94241    6.75865   1.175 0.240303
## PersonalM:DivSepMar   -13.30743    8.70654  -1.528 0.126818
## PersonalM:Single       30.06153    4.33588   6.933 8.78e-12 ***
## ForeignYes            -54.42600    9.65159  -5.639 2.41e-08 ***
## PlansNone              32.27987    5.16452   6.250 6.81e-10 ***
## PlansStores            11.89419    9.68139   1.229 0.219615
## OtherPartiesGuarantor  61.12648   12.15645   5.028 6.17e-07 ***
## OtherPartiesNone       11.94918    8.90652   1.342 0.180118
## Disposable2           -11.97283    6.48331  -1.847 0.065176 .
## Disposable3           -26.72090    6.94964  -3.845 0.000131 ***
## Disposable4           -41.89259    6.66438  -6.286 5.48e-10 ***
## Amount                -23.88738    3.73000  -6.404 2.64e-10 ***
## TelephoneYes           12.64238    3.87933   3.259 0.001168 **
## HousingRent           -16.91325    4.89081  -3.458 0.000574 ***
## HousingRentFree        -3.76808    6.15605  -0.612 0.540658
## Age                    16.27202    6.74177   2.414 0.016030 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.4 on 762 degrees of freedom
## Multiple R-squared:  0.7181, Adjusted R-squared:  0.7044
## F-statistic: 52.46 on 37 and 762 DF,  p-value: < 2.2e-16
```

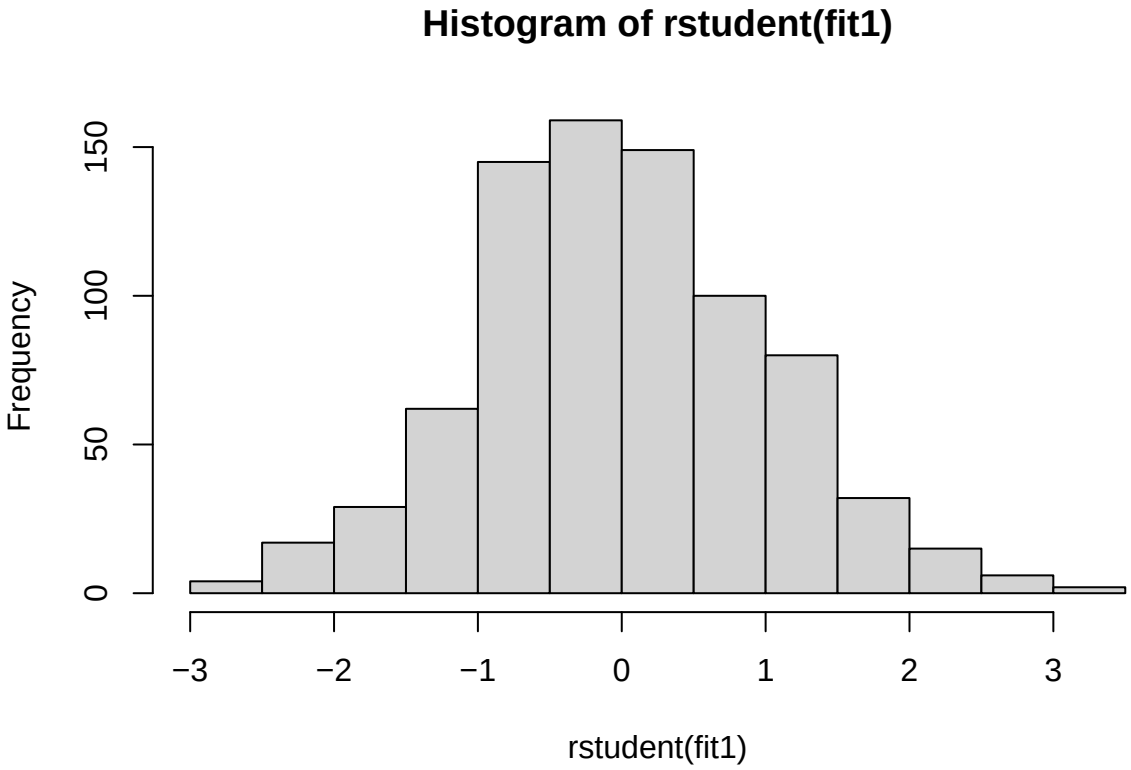
We see that the model is the same as *fit1*, so we just keep *fit1* as our chosen best model.

Diagnostic checks

Let's check whether error is normally distributed.



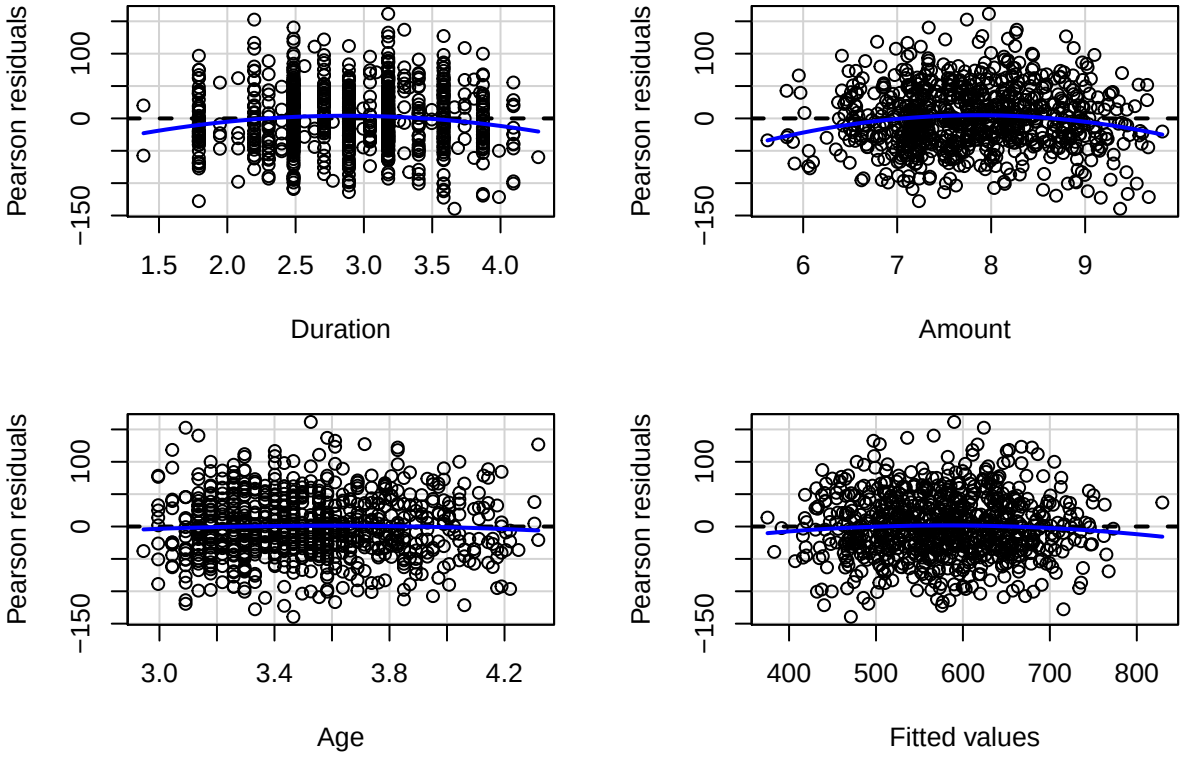

```
## [1] 272 639
```

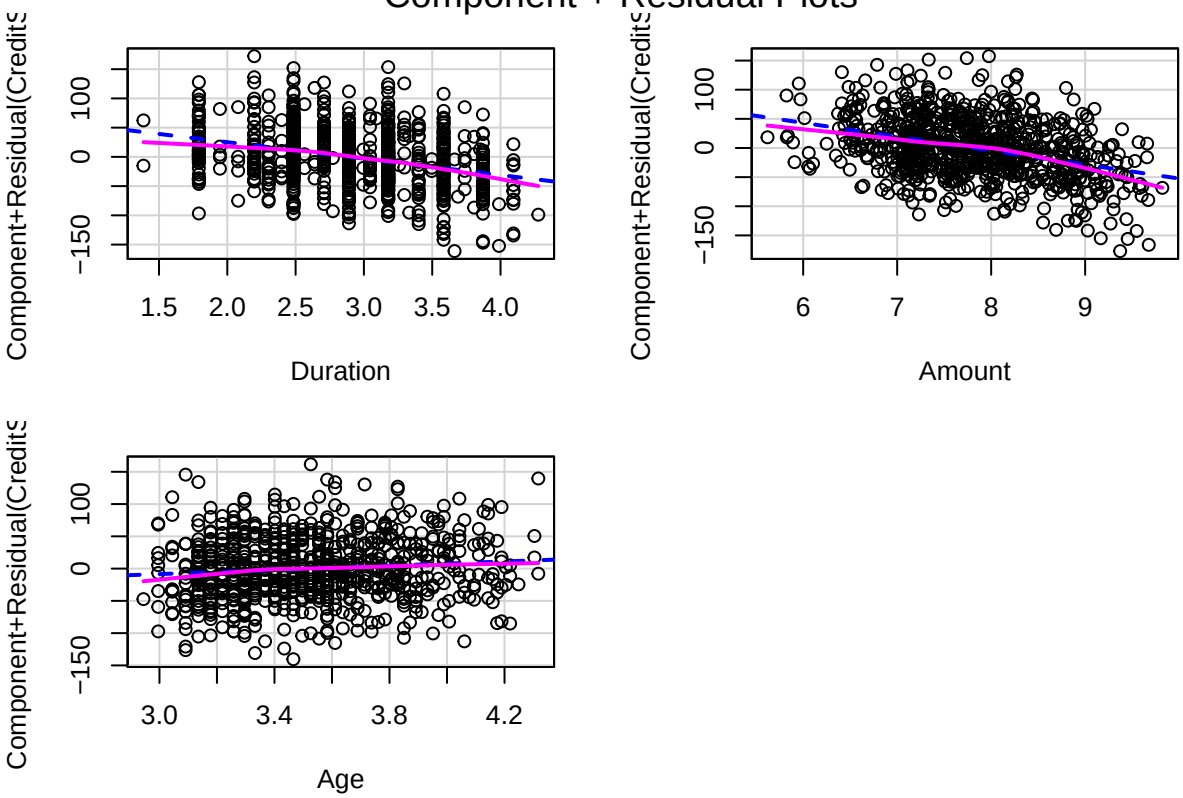


Interpretation:

Although the QQ-plot of *fit1* is not completely a straight line, almost all points are inside the confidence interval and the histogram of the residuals is approximately normal distributed. Therefore, the error is approximately normal distributed and the small skewness will not cause much affect to the results. Note that the two outlying data points has been highlighted, *no.*272 and *no.*639, considering the skewness might be caused by outliers, we want to know which in order to see what happens if they are removed later.

Let's check whether errors have constant variances.





Interpretation:

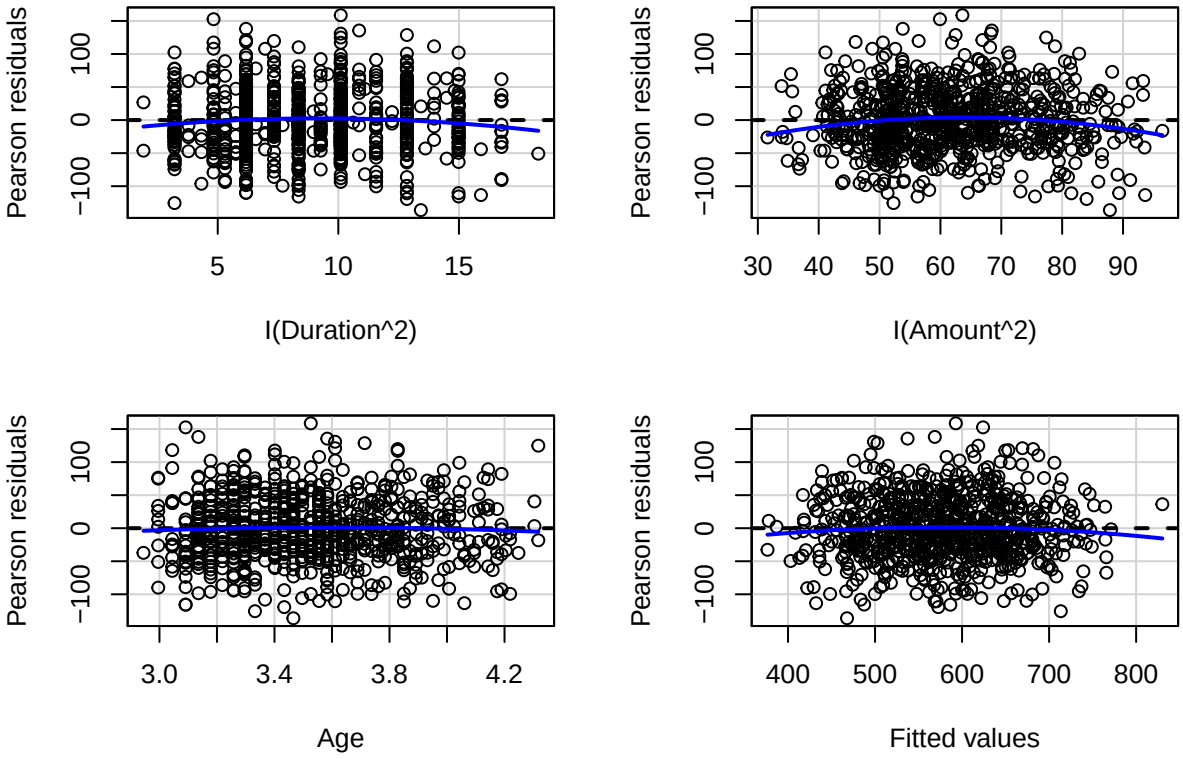
The residual plot for $log(Age)$ is approximately straight line hence the skewness of residual plot for age might

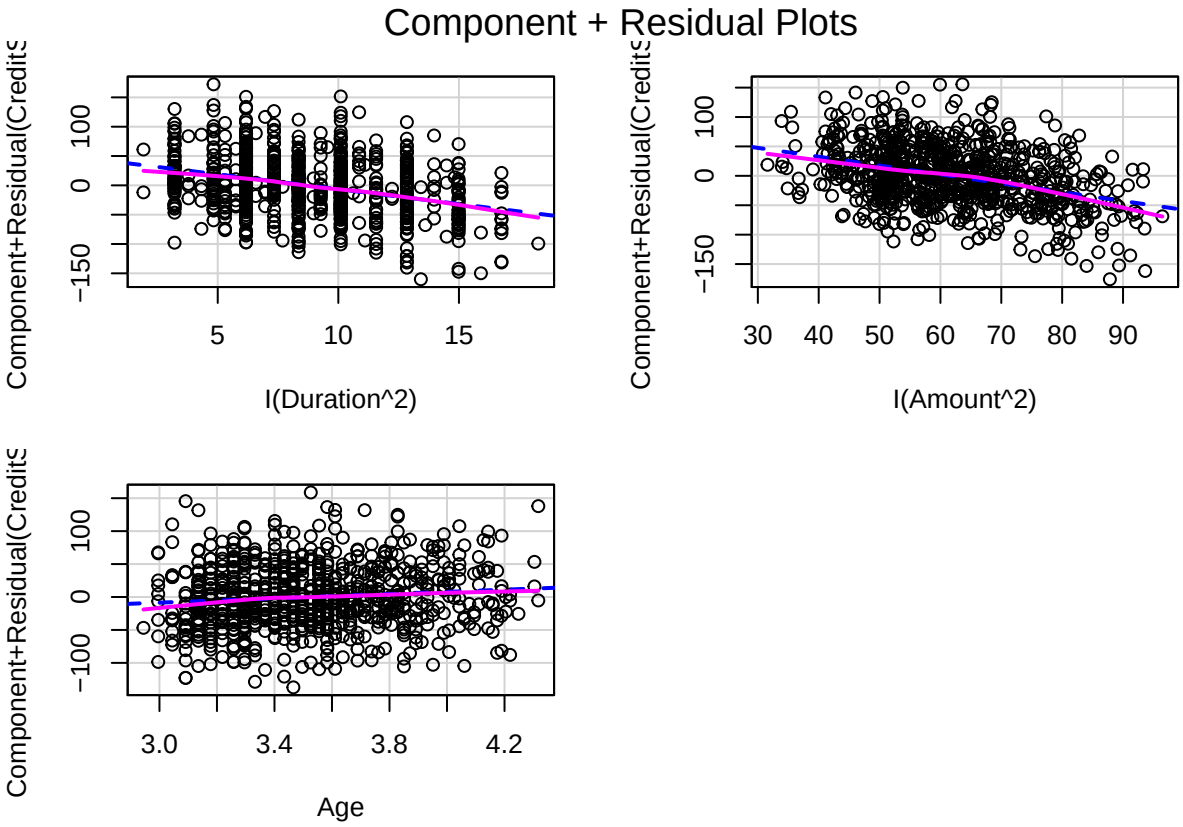
be caused by outliers, while for $log(Duration)$ and $log(Amount)$, plots tends to be downward so the trend is a little bit non-monotonic but it is not too bad. So we just try power transformation for them to see what happen, called model *fit2*.

$log(Duration) \rightarrow (log(Duration))^2$

$log(Amount) \rightarrow (log(Amount))^2$

Let's check the residual plots as well.





Interpretation:

We see that the plots become much better and all the Component+Residual Plots are approximately straight lines. Moreover, the residual plots are approximately around 0 and have no clear patterns except some outliers, which is acceptable. Thus, the assumptions for linear regression are satisfied.

So we keep *fit2* as our best model and let's check the summary output.

```
##
## Call:
## lm(formula = CreditScore ~ Status + I(Duration^2) + History +
##     Purpose + I(Amount^2) + Savings + Disposable + Personal +
##     OtherParties + Age + Plans + Housing + Telephone + Foreign,
##     data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -136.226  -30.894   -2.888   32.271  158.669
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          642.67279   33.25874  19.323  < 2e-16 ***
## StatusNegative       -50.33865    8.09329  -6.220 8.20e-10 ***
## StatusNone            37.80254    7.87905   4.798 1.93e-06 ***
## StatusSmall          -30.63848    8.19473  -3.739 0.000199 ***
## I(Duration^2)         -5.05929    0.79482  -6.365 3.36e-10 ***
## HistoryB              -1.71064   12.30882  -0.139 0.889505
## HistoryC              47.43691    9.41143   5.040 5.81e-07 ***
## HistoryD              48.61100   10.73529   4.528 6.90e-06 ***
```

```
## HistoryE               80.96960    9.74846   8.306 4.51e-16 ***
## PurposeDomestic       -24.04207   17.00420  -1.414 0.157803
## PurposeEducation      -53.05132   10.09636  -5.254 1.93e-07 ***
## PurposeFurniture       -0.65449    7.33330  -0.089 0.928907
## PurposeNewCar         -46.39522    7.00577  -6.622 6.67e-11 ***
## PurposeOther          -17.52377   16.49235  -1.063 0.288328
## PurposeRepairs        -26.72275   13.38743  -1.996 0.046278 *
## PurposeTelevision       2.33951    6.89109   0.339 0.734328
## PurposeTraining        60.86332   21.52967   2.827 0.004823 **
## PurposeUsedCar         37.63425    8.18310   4.599 4.97e-06 ***
## I(Amount^2)            -1.50658    0.23836  -6.321 4.43e-10 ***
## SavingsLow            -13.21931    7.81988  -1.690 0.091346 .
## SavingsMedium           0.09274    9.23287   0.010 0.991988
## SavingsUnknown         25.96073    8.54548   3.038 0.002463 **
## SavingsVeryLarge       61.98826   11.23031   5.520 4.66e-08 ***
## Disposable2           -12.53188    6.41867  -1.952 0.051255 .
## Disposable3           -27.68162    6.88491  -4.021 6.38e-05 ***
## Disposable4           -42.74550    6.56140  -6.515 1.32e-10 ***
## PersonalF:Single        7.66456    6.71791   1.141 0.254264
## PersonalM:DivSepMar   -13.69298    8.65364  -1.582 0.113988
## PersonalM:Single       29.91992    4.30413   6.951 7.78e-12 ***
## OtherPartiesGuarantor  61.43833   12.08982   5.082 4.71e-07 ***
## OtherPartiesNone       12.31841    8.85811   1.391 0.164742
## Age                    16.45256    6.69879   2.456 0.014270 *
## PlansNone              32.60000    5.13389   6.350 3.70e-10 ***
## PlansStores            12.05621    9.62299   1.253 0.210643
## HousingRent           -17.08724    4.86161  -3.515 0.000466 ***
## HousingRentFree        -2.90474    6.12398  -0.474 0.635406
## TelephoneYes           12.83856    3.86178   3.325 0.000928 ***
## ForeignYes            -54.70072    9.58834  -5.705 1.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.11 on 762 degrees of freedom
## Multiple R-squared:  0.7214, Adjusted R-squared:  0.7078
## F-statistic: 53.32 on 37 and 762 DF,  p-value: < 2.2e-16
```

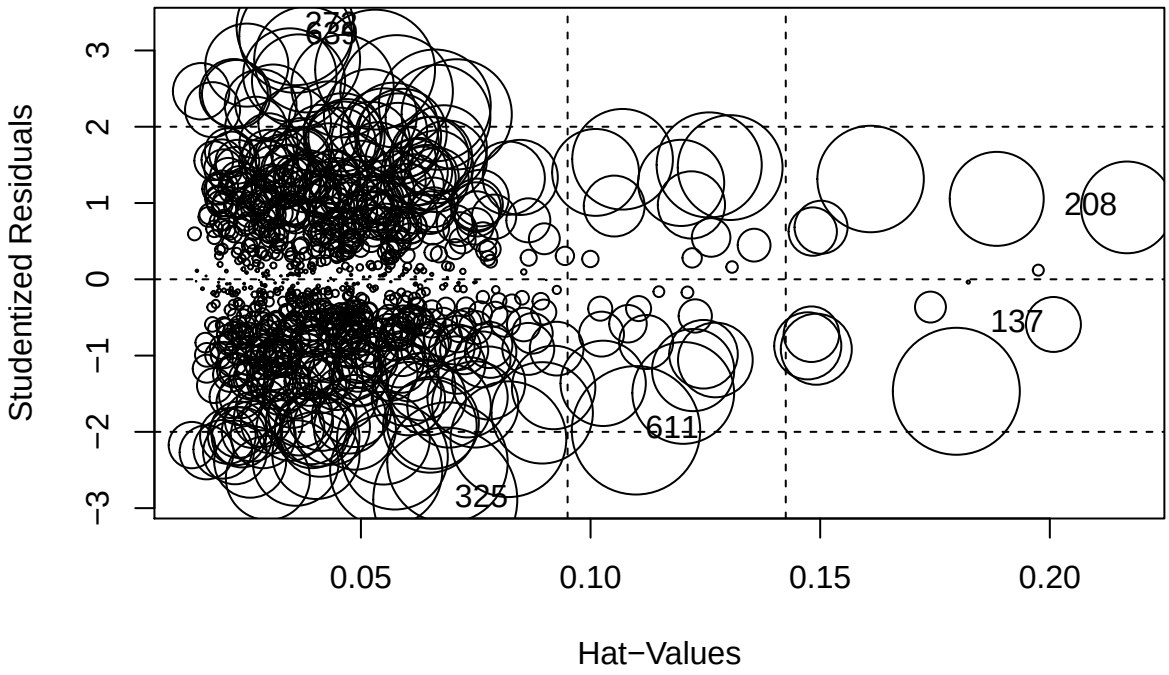
```
## [1] "fit2$adj.r.squared = 0.707831"
```

Interpretation:

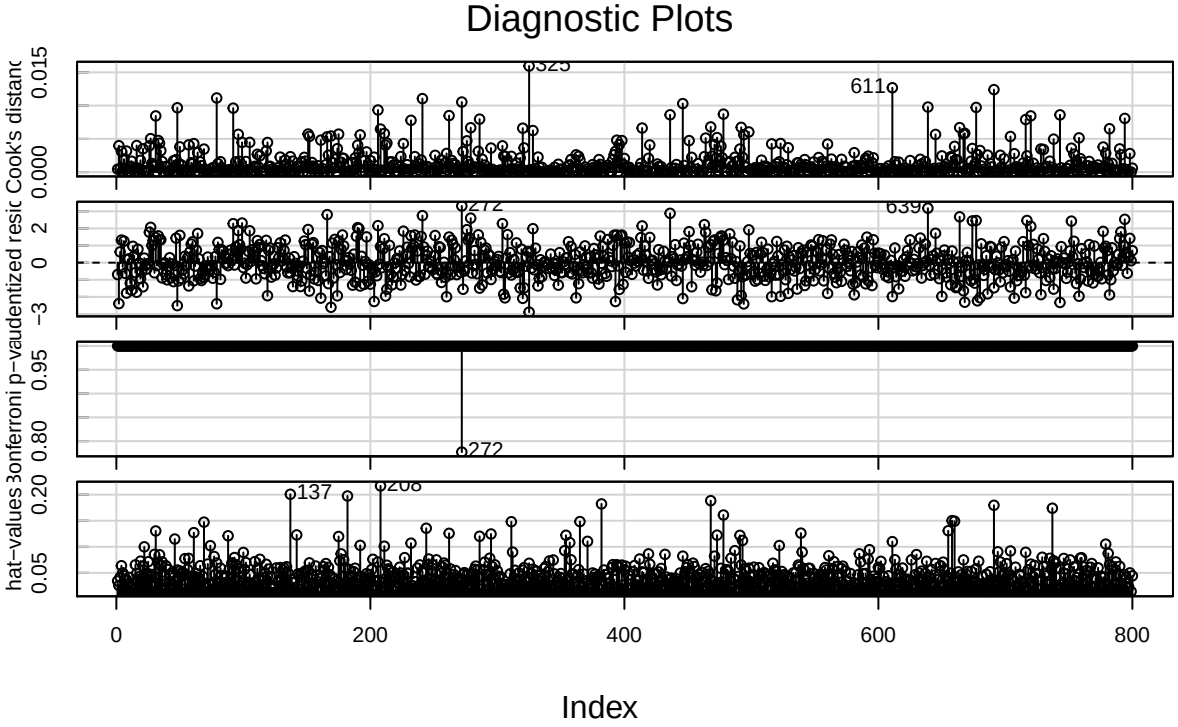
We find that $(log(Duration))^2$ and $(log(Amount)^2$ are still significant at 5% level and R^2_{adj} has been increased to 0.707831, which shows a good fit of our model *fit2*.

Moreover, we also need to test whether there are unusual or influential observations which could affect the result.

Let's look for outliers and high-leverage points.




```
##         StudRes        Hat       CookD
## 137 -0.5921394 0.20079092 0.002320161
## 208  0.9425215 0.21679834 0.006472095
## 272  3.3116139 0.03566543 0.010535907
## 325 -2.8877540 0.06836724 0.015950606
## 611 -1.9796145 0.10988883 0.012683133
## 639  3.1836112 0.03579505 0.009784403
```



Interpretation:

From the plots and the output values:

Observations *no.*272,*no.*639,*no.*325 have a high value of |StudRes| > 2 so these three observations are considered as outliers.

Observation *no.*137 and *no.*208 have high Hat values much larger than $2p/n = 15 \times 2/800 = 0.0375$ so considered as high-leverage points. However, the CookD of them are very small so these two points are not influential.

Observation *no.*325 and *no.*611 have a high value of CookD, so these two observations could be influential.

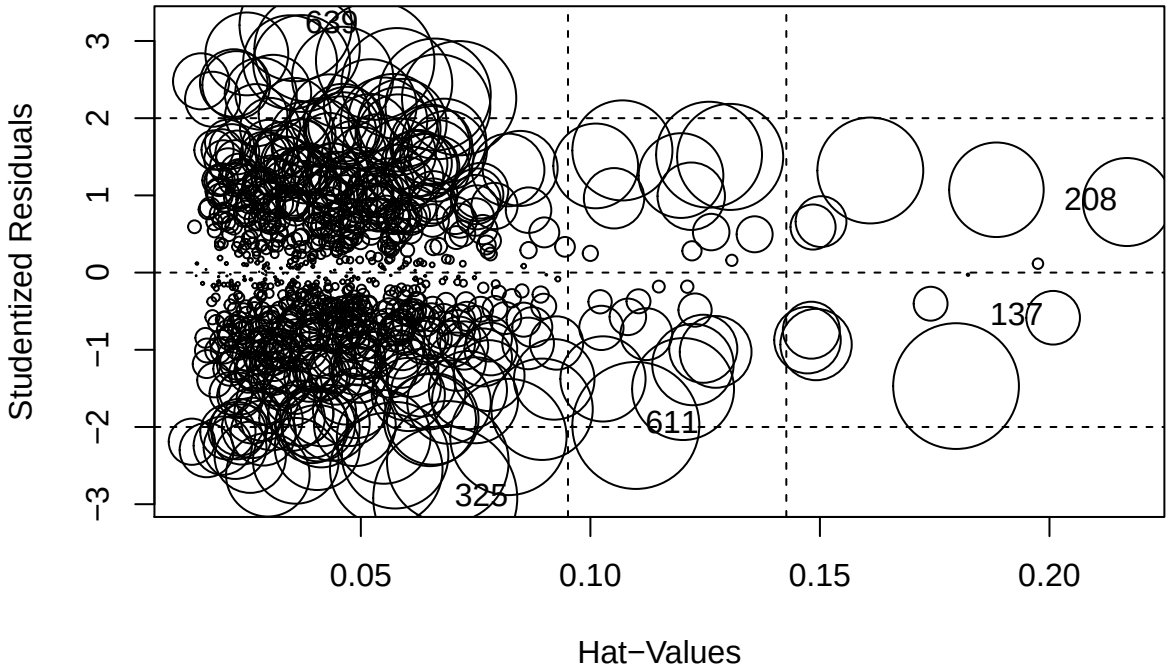
However, the value of CookD don't differ so much and all of them are < 1 so we focus on the StudRes and Hat-Values first.

Let's consider the model with the removal of points:

fit3: model *fit2* with the removal of observation *no.272*, which has the highest value of |StudRes|.

```
## Calls:
## 1: lm(formula = CreditScore ~ Status + I(Duration^2) + History + Purpose +
##   I(Amount^2) + Savings + Disposable + Personal + OtherParties + Age + Plans
##   + Housing + Telephone + Foreign, data = Train)
## 2: lm(formula = fit2, data = Train1)
##
##                        Model 1 Model 2
## (Intercept)               643     639
## StatusNegative          -50.3   -51.2
## StatusNone               37.8    37.7
## StatusSmall             -30.6   -30.9
## I(Duration^2)           -5.06   -5.13
## HistoryB                -1.71   -1.55
## HistoryC                 47.4    47.0
## HistoryD                 48.6    48.5
## HistoryE                   81      81
## PurposeDomestic         -24.0   -23.4
## PurposeEducation        -53.1   -52.4
## PurposeFurniture       -0.654  -0.361
## PurposeNewCar           -46.4   -46.1
## PurposeOther            -17.5   -17.1
## PurposeRepairs          -26.7   -26.4
## PurposeTelevision        2.34    2.43
## PurposeTraining          60.9    60.6
## PurposeUsedCar           37.6    35.8
## I(Amount^2)             -1.51   -1.46
## SavingsLow              -13.2   -13.4
## SavingsMedium          0.0927  0.0765
## SavingsUnknown           26.0    26.1
## SavingsVeryLarge           62      62
## Disposable2             -12.5   -13.3
## Disposable3             -27.7   -27.2
## Disposable4             -42.7   -42.3
## PersonalF:Single         7.66    7.82
## PersonalM:DivSepMar     -13.7   -13.8
## PersonalM:Single         29.9    29.6
## OtherPartiesGuarantor    61.4    61.7
## OtherPartiesNone         12.3    12.3
## Age                      16.5    17.1
```

```
## PlansNone                32.6    32.3
## PlansStores              12.1    12.1
## HousingRent             -17.1   -16.8
## HousingRentFree          -2.9    -4.6
## TelephoneYes             12.8    12.2
## ForeignYes              -54.7   -54.6
```




```
##        StudRes        Hat       CookD
## 137 -0.5848924 0.20079992 0.002263873
## 208  0.9141369 0.21688358 0.006091623
## 325 -2.9202828 0.06838213 0.016311577
## 611 -1.9798528 0.10990184 0.012687795
## 639  3.2050902 0.03579506 0.009914958
```

```
## [1] "fit3$adj.r.squared = 0.710646"
```

Interpretation:

After we delete *no.*272,the CookD for the whole model changed but *no.*639 still has the largest high |StudtRes|. Notice that R^2_{adj} has been increased to 0.710646. Let's try removing *no.*639 to see what happens, called *fit4*.

```
## Calls:
## 1: lm(formula = CreditScore ~ Status + I(Duration^2) + History + Purpose +
##    I(Amount^2) + Savings + Disposable + Personal + OtherParties + Age + Plans
##    + Housing + Telephone + Foreign, data = Train)
## 2: lm(formula = fit2, data = Train1)
## 3: lm(formula = fit2, data = Train2)
##
##                       Model 1 Model 2 Model 3
## (Intercept)               643     639     639
## StatusNegative          -50.3   -51.2   -51.3
## StatusNone               37.8    37.7    37.1
## StatusSmall             -30.6   -30.9   -30.7
## I(Duration^2)           -5.06   -5.13   -5.00
## HistoryB                -1.71   -1.55   -1.61
```

```
## HistoryC                47.4    47.0    46.8
## HistoryD                48.6    48.5    48.4
## HistoryE                  81      81      81
## PurposeDomestic        -24.0   -23.4   -23.2
## PurposeEducation       -53.1   -52.4   -52.0
## PurposeFurniture      -0.654  -0.361  -1.242
## PurposeNewCar          -46.4   -46.1   -45.9
## PurposeOther           -17.5   -17.1   -17.6
## PurposeRepairs         -26.7   -26.4   -26.2
## PurposeTelevision       2.34    2.43    2.51
## PurposeTraining         60.9    60.6    60.6
## PurposeUsedCar          37.6    35.8    36.1
## I(Amount^2)            -1.51   -1.46   -1.48
## SavingsLow             -13.2   -13.4   -13.4
## SavingsMedium         0.0927  0.0765 -2.0019
## SavingsUnknown          26.0    26.1    26.1
## SavingsVeryLarge        62.0    62.0    61.9
## Disposable2            -12.5   -13.3   -14.4
## Disposable3            -27.7   -27.2   -27.5
## Disposable4            -42.7   -42.3   -42.8
## PersonalF:Single        7.66    7.82    8.35
## PersonalM:DivSepMar    -13.7   -13.8   -13.2
## PersonalM:Single        29.9    29.6    30.0
## OtherPartiesGuarantor   61.4    61.7    61.1
## OtherPartiesNone        12.3    12.3    11.6
## Age                     16.5    17.1    17.5
## PlansNone               32.6    32.3    32.2
## PlansStores             12.1    12.1    12.0
## HousingRent            -17.1   -16.8   -17.7
## HousingRentFree        -2.90   -4.60   -4.93
## TelephoneYes            12.8    12.2    12.5
## ForeignYes             -54.7   -54.6   -54.9
```

```
##
## Call:
## lm(formula = fit2, data = Train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.649  -30.154   -2.857   32.584  138.064
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           639.4449    32.8638  19.457  < 2e-16 ***
## StatusNegative        -51.2620     7.9966  -6.410 2.55e-10 ***
## StatusNone             37.1499     7.7828   4.773 2.17e-06 ***
## StatusSmall           -30.6810     8.0930  -3.791 0.000162 ***
## I(Duration^2)          -5.0032     0.7861  -6.364 3.39e-10 ***
## HistoryB               -1.6137    12.1554  -0.133 0.894422
## HistoryC               46.7695     9.2952   5.032 6.08e-07 ***
## HistoryD               48.3784    10.6015   4.563 5.87e-06 ***
## HistoryE               81.0095     9.6269   8.415  < 2e-16 ***
## PurposeDomestic       -23.1777    16.7933  -1.380 0.167939
## PurposeEducation      -52.0207     9.9731  -5.216 2.36e-07 ***
```

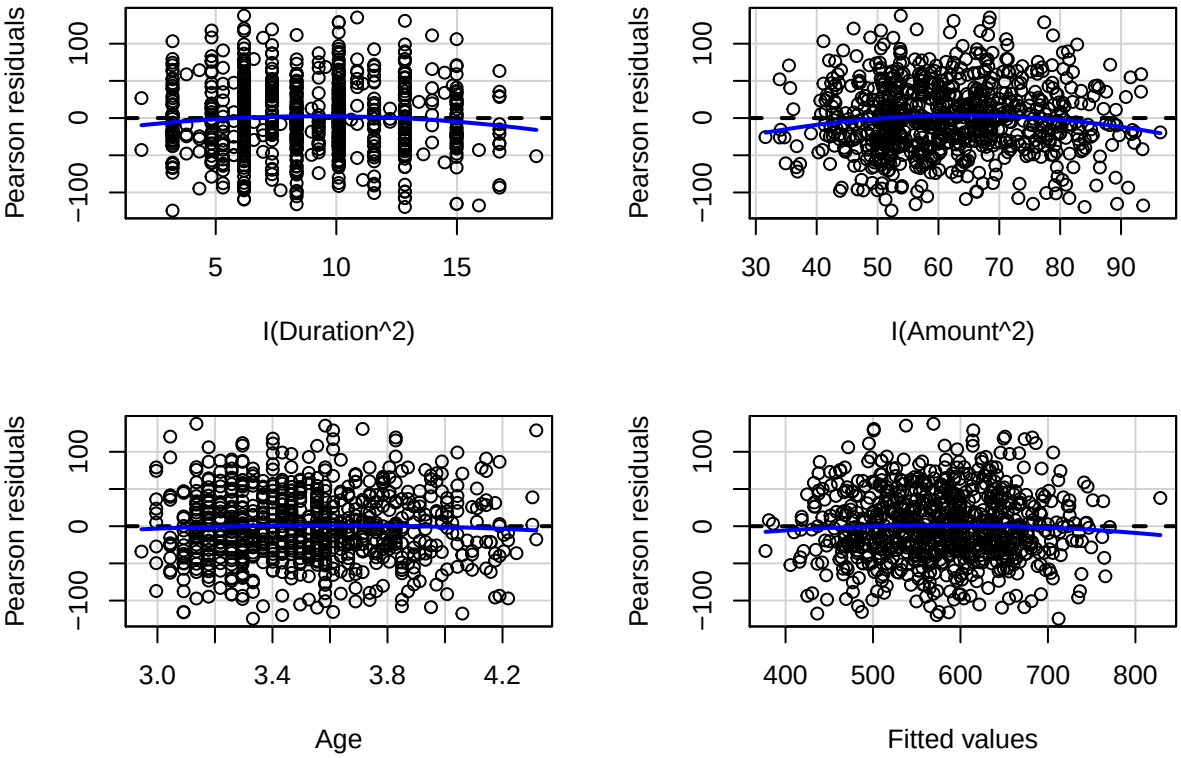
```
## PurposeFurniture       -1.2417     7.2476  -0.171 0.864012
## PurposeNewCar         -45.9375     6.9191  -6.639 6.00e-11 ***
## PurposeOther          -17.6155    16.2881  -1.081 0.279819
## PurposeRepairs        -26.1998    13.2209  -1.982 0.047874 *
## PurposeTelevision       2.5112     6.8052   0.369 0.712223
## PurposeTraining        60.5980    21.2613   2.850 0.004488 **
## PurposeUsedCar         36.0965     8.0996   4.457 9.58e-06 ***
## I(Amount^2)            -1.4762     0.2359  -6.258 6.53e-10 ***
## SavingsLow            -13.4453     7.7226  -1.741 0.082081 .
## SavingsMedium          -2.0019     9.1407  -0.219 0.826699
## SavingsUnknown         26.0741     8.4391   3.090 0.002077 **
## SavingsVeryLarge       61.8698    11.0903   5.579 3.37e-08 ***
## Disposable2           -14.4119     6.3525  -2.269 0.023565 *
## Disposable3           -27.4673     6.8014  -4.038 5.93e-05 ***
## Disposable4           -42.7765     6.4829  -6.598 7.79e-11 ***
## PersonalF:Single        8.3528     6.6363   1.259 0.208546
## PersonalM:DivSepMar   -13.2010     8.5475  -1.544 0.122902
## PersonalM:Single       29.9837     4.2540   7.048 4.07e-12 ***
## OtherPartiesGuarantor  61.1499    11.9402   5.121 3.85e-07 ***
## OtherPartiesNone       11.6303     8.7502   1.329 0.184202
## Age                    17.4768     6.6191   2.640 0.008452 **
## PlansNone              32.1997     5.0706   6.350 3.70e-10 ***
## PlansStores            11.9716     9.5030   1.260 0.208137
## HousingRent           -17.6743     4.8084  -3.676 0.000254 ***
## HousingRentFree        -4.9283     6.0699  -0.812 0.417094
## TelephoneYes           12.5080     3.8186   3.276 0.001103 **
## ForeignYes            -54.8567     9.4691  -5.793 1.01e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.5 on 760 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7128
## F-statistic: 54.45 on 37 and 760 DF,  p-value: < 2.2e-16
```

```
## [1] "fit4$adj.r.squared = 0.712758"
```

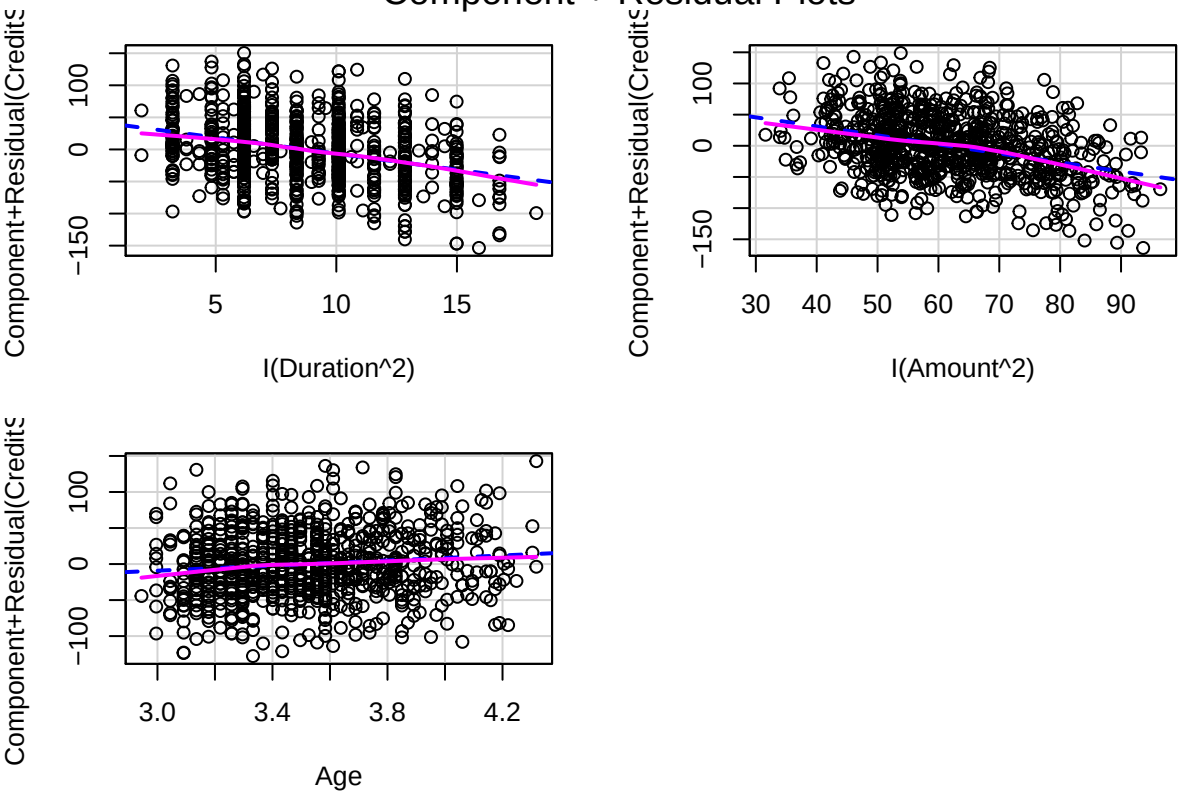
Interpretation:

According to the comparison and summary, the model changed after deleting the observation *no.272* and *no.639*. Coefficient of $log(Age)$ has been increased by about 6% and $log(Age)$ become more significant by summary output. The R^2_{adj} has been increased to 0.712758.

Let's remove *no.*325 to see what happens, called *fit5*.



Component + Residual Plots




```
## Calls:
## 1: lm(formula = CreditScore ~ Status + I(Duration^2) + History + Purpose +
##    I(Amount^2) + Savings + Disposable + Personal + OtherParties + Age + Plans
##    + Housing + Telephone + Foreign, data = Train)
```

```
## 2: lm(formula = fit2, data = Train1)
## 3: lm(formula = fit2, data = Train2)
## 4: lm(formula = fit2, data = Train3)
##
##                       Model 1 Model 2 Model 3 Model 4
## (Intercept)               643     639     639     635
## StatusNegative          -50.3   -51.2   -51.3   -51.2
## StatusNone               37.8    37.7    37.1    37.0
## StatusSmall             -30.6   -30.9   -30.7   -30.3
## I(Duration^2)           -5.06   -5.13   -5.00   -5.01
## HistoryB                -1.71   -1.55   -1.61   -1.48
## HistoryC                 47.4    47.0    46.8    46.7
## HistoryD                 48.6    48.5    48.4    50.4
## HistoryE                   81      81      81      81
## PurposeDomestic         -24.0   -23.4   -23.2   -22.0
## PurposeEducation        -53.1   -52.4   -52.0   -47.5
## PurposeFurniture       -0.654  -0.361  -1.242  -0.473
## PurposeNewCar           -46.4   -46.1   -45.9   -45.3
## PurposeOther            -17.5   -17.1   -17.6   -17.4
## PurposeRepairs          -26.7   -26.4   -26.2   -25.7
## PurposeTelevision        2.34    2.43    2.51    3.41
## PurposeTraining          60.9    60.6    60.6    62.4
## PurposeUsedCar           37.6    35.8    36.1    36.5
## I(Amount^2)             -1.51   -1.46   -1.48   -1.45
## SavingsLow              -13.2   -13.4   -13.4   -13.7
## SavingsMedium          0.0927  0.0765 -2.0019 -0.7210
## SavingsUnknown           26.0    26.1    26.1    25.6
## SavingsVeryLarge         62.0    62.0    61.9    61.8
## Disposable2             -12.5   -13.3   -14.4   -13.7
## Disposable3             -27.7   -27.2   -27.5   -27.3
## Disposable4             -42.7   -42.3   -42.8   -42.5
## PersonalF:Single         7.66    7.82    8.35    8.49
## PersonalM:DivSepMar     -13.7   -13.8   -13.2   -13.1
## PersonalM:Single         29.9    29.6    30.0    30.3
## OtherPartiesGuarantor    61.4    61.7    61.1    61.4
## OtherPartiesNone         12.3    12.3    11.6    11.9
## Age                      16.5    17.1    17.5    17.7
## PlansNone                32.6    32.3    32.2    32.4
## PlansStores              12.1    12.1    12.0    11.8
## HousingRent             -17.1   -16.8   -17.7   -16.6
## HousingRentFree         -2.90   -4.60   -4.93   -5.57
## TelephoneYes             12.8    12.2    12.5    12.8
## ForeignYes              -54.7   -54.6   -54.9   -55.2

## [1] "fit2$adj.r.squared = 0.707831"

## [1] "fit3$adj.r.squared = 0.710646"

## [1] "fit4$adj.r.squared = 0.712758"

## [1] "fit5$adj.r.squared = 0.713202"
```

Interpretation:

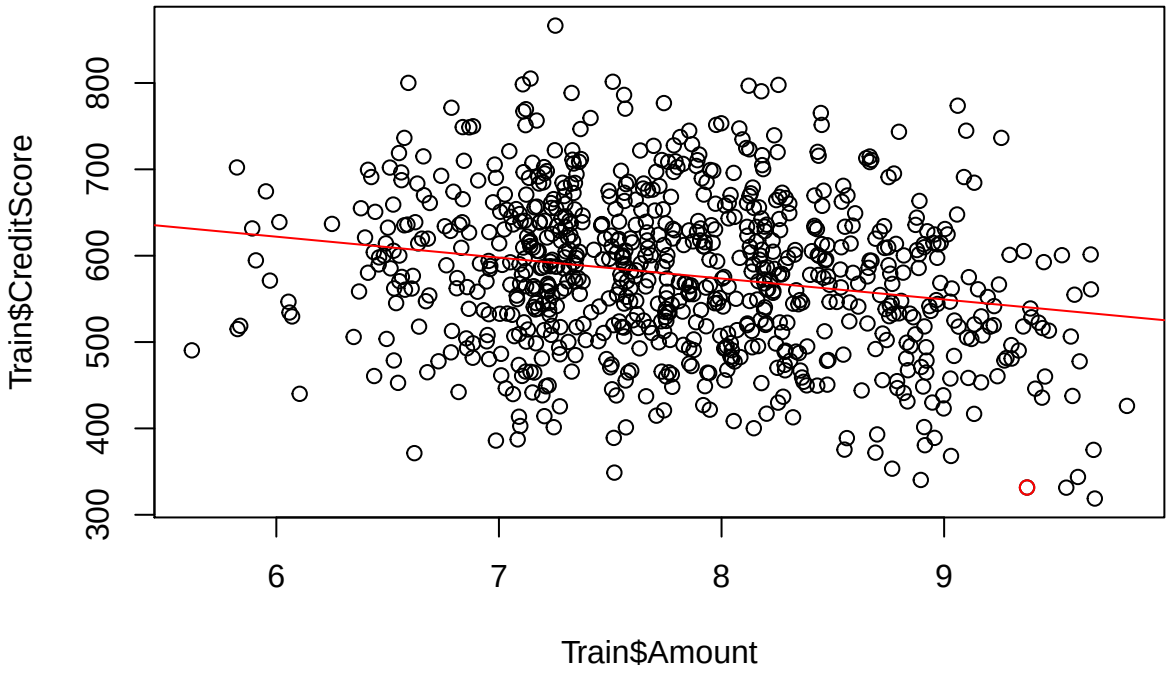
After deleting *no.272,no.639* and *no.325*, according to the comparison and summary: $log(Age)$ become more significant by summary output. The R^2_{adj} has been increased to 0.713202. However, the differences are very small so this is further evidence that it is not a clear call, and it is not right or wrong to remove these

observations. Depending on what we want to do with the model (e.g. if we want to make predictions), we could even use both models and see how the outcome differs. According to residual and Component+Residual Plot, the assumption still holds.

If we take a look at the *no.*325, we find that it has higher Amount but lower CreditScore extremely (the red points shown below). So this observation should be deleted since it may influence the accuracy of prediction for general individuals.

Actually, it is hard to say whether we should remove these three observations, we should be very careful when considering the deletion of observations. This could be discussed separately. We can determine that the observation is an outlier because of data errors in recording or because a protocal wasn't followed. In these cases, deleting the offending observation seems reasonable. However, some observations may have some interesting characteristics, which might give useful insight to our study, which is a subjective judgment. Therefore, we evaluate them in prediction part.

```
##     Status Duration History   Purpose   Amount Savings Employment Disposable
## 325  Small 3.663562       D Education 9.372459  Medium       Long          2
##     Personal OtherParties Residence Property      Age Plans Housing Existing
## 325 M:Single         None         3     None 3.465736  None    Rent        1
##         Job Dependants Telephone Foreign CreditScore
## 325 Skilled          1       Yes     Yes    331.5908
```



Model Prediction

Summary of current model:

fit2: best model we find:

$CreditScore \sim Status + (log(Duration))^2 + History + Purpose + (log(Amount))^2 + Savings + Disposable + Personal + OtherParties + log(Age) + Plans + Housing + Telephone + Foreign$

fit3: best model with the removal of *no.272*.

fit4: best model with the removal of both *no.272* and *no.639*.

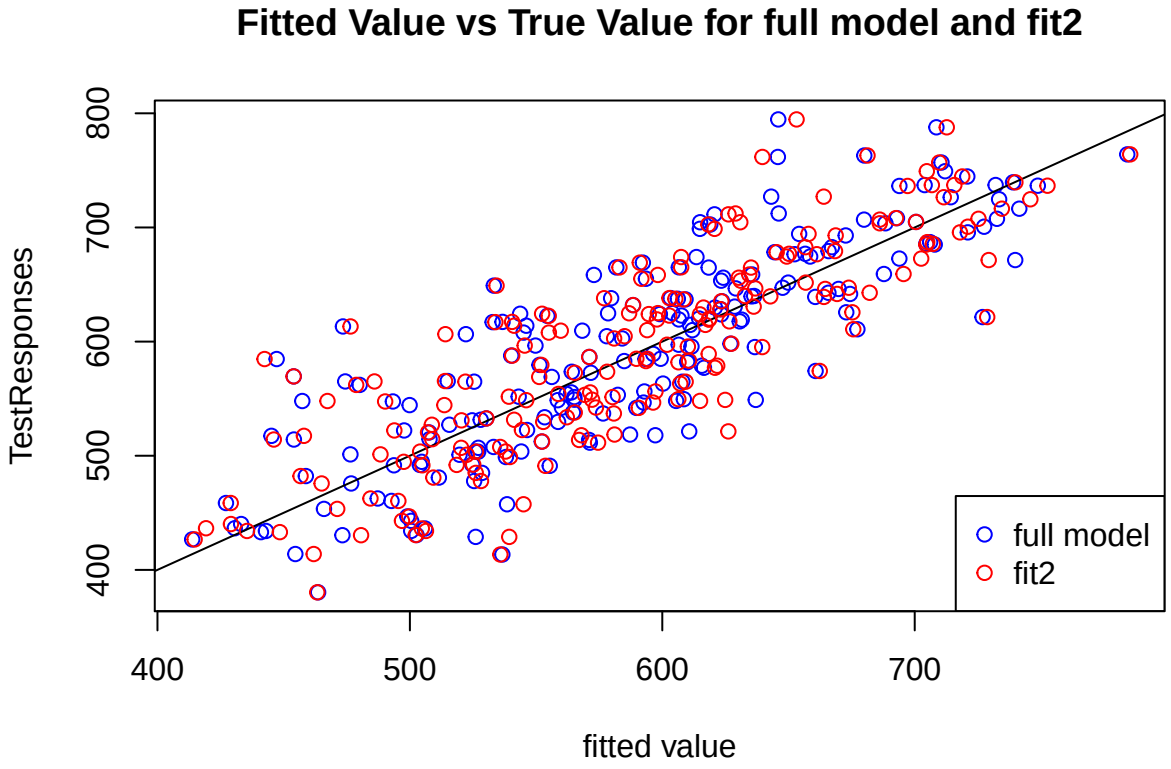
fit5: best model with the removal of both *no.272*,*no.639* and *no.325*.

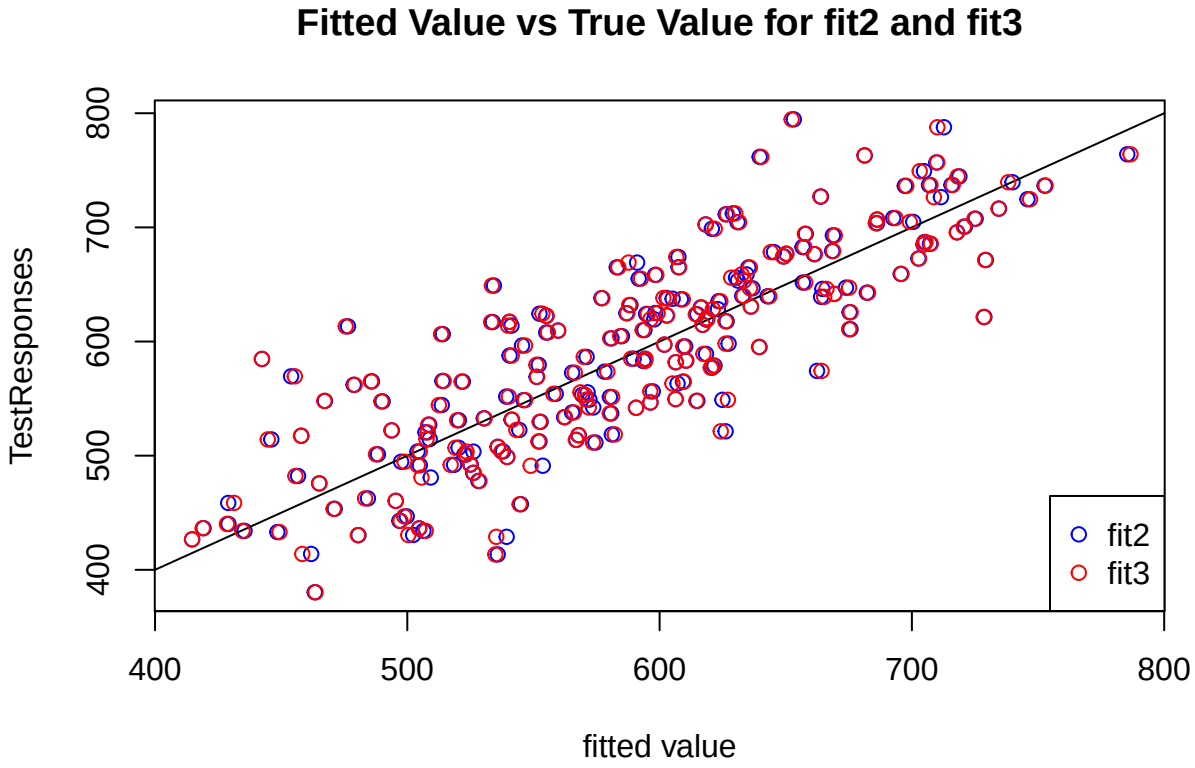
We will use the *Test.txt* data to predict the responses for individuals.

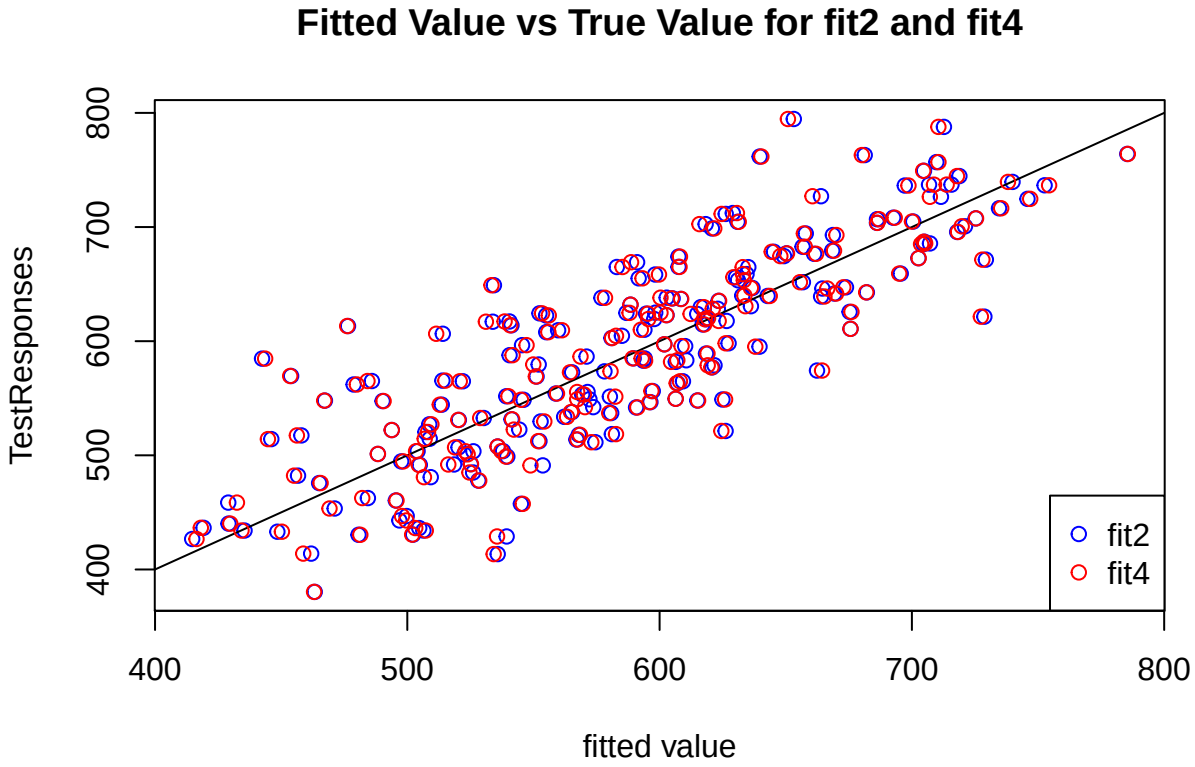
We use the mean-square error (MSE) as a criterion of the quality for *full model*,*fit2*, *fit3*, *fit4* and *fit5*.

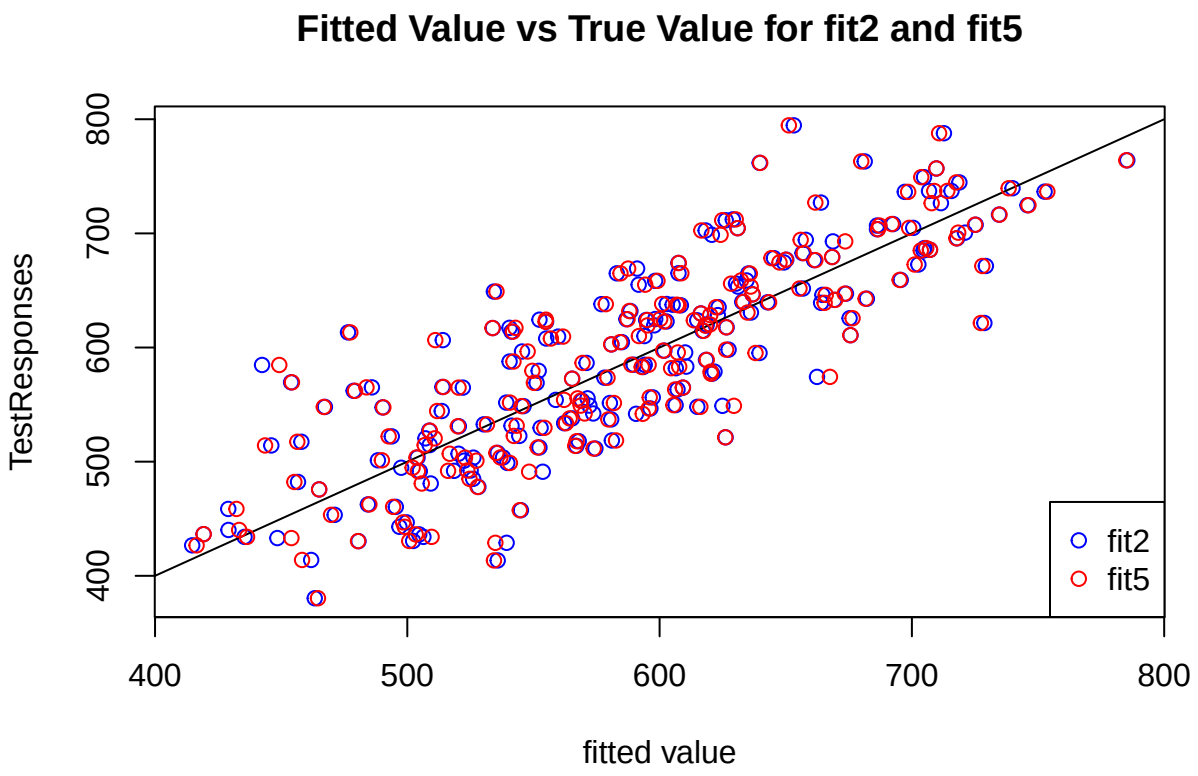
```
## [1] "full_MSE = 2468.280381"
## [1] "fit2_MSE = 2443.476907"
## [1] "fit3_MSE = 2435.199877"
## [1] "fit4_MSE = 2436.308412"
## [1] "fit5_MSE = 2425.233652"
```

Let's also draw the graph for fitted value and True Value for models.




Fitted Value vs True Value for fit2 and fit3
TestResponses
400
500
600
700
800
fitted value
fit2
fit3


Fitted Value vs True Value for fit2 and fit4
TestResponses
400
500
600
700
800
fitted value
fit2
fit4



Interpretation:

Compared to full model, our best model *fit2* has a lower MSE 2443.477. The plot shows that *fit2* is much better than *full model* 2468.28 since the points concentrate more along the $y = x$ axis.

We see that the MSE tends to decreases as we delete the outliers and *fit5* has the lowest MSE 2425.234. So our best model is now *fit5*.

Let's summarize our data into a data frame and predict:

```
##     fitted_value lower_bound upper_bound true_value
## 1       477.4090    380.6238    574.1942   613.2571
## 2       695.1527    596.9392    793.3662   659.2217
## 3       588.8279    492.6674    684.9884   584.9637
## 4       643.4013    547.7542    739.0484   639.6288
## 5       584.3633    486.9702    681.7564   604.7193
## 6       619.9772    522.5292    717.4252   620.1996
## 7       533.7640    437.5979    629.9301   617.0603
## 8       562.1437    463.9633    660.3240   554.1421
## 9       596.0893    500.3294    691.8491   556.3166
## 10      606.8638    509.9590    703.7686   637.4851
## 11      708.8831    611.7069    806.0593   737.1314
## 12      500.7662    402.9024    598.6300   430.5130
## 13      490.5075    394.8866    586.1284   547.5392
## 14      657.1201    560.8720    753.3682   682.7224
## 15      650.0193    552.3720    747.6665   676.9100
## 16      701.0991    603.3159    798.8823   672.7252
## 17      628.3286    532.1093    724.5479   655.9837
## 18      636.0935    538.8844    733.3026   653.3876
## 19      624.8229    528.6831    720.9628   711.4867
## 20      717.5793    617.9138    817.2447   744.6560
```

```
## 21      705.6903    603.9975    807.3831   687.1201
## 22      703.5347    606.5239    800.5456   685.0312
## 23      543.3457    446.5268    640.1645   531.5324
## 24      687.1644    589.0400    785.2888   706.8683
## 25      489.9192    391.2941    588.5444   501.2033
## 26      479.2898    382.2074    576.3722   562.0501
## 27      522.8847    424.1405    621.6288   503.5416
## 28      634.8224    538.3258    731.3190   630.6475
## 29      587.5301    490.6970    684.3632   669.1843
## 30      509.6301    412.0774    607.1828   434.1115
## 31      673.5630    575.9498    771.1762   692.9877
## 32      573.8576    470.1270    677.5883   511.4866
## 33      681.6501    585.9872    777.3129   642.7761
## 34      746.2124    650.0412    842.3836   724.6790
## 35      514.2817    414.7035    613.8599   565.4501
## 36      727.3414    631.4962    823.1865   621.4969
## 37      607.5850    510.0289    705.1411   674.0797
## 38      607.5529    510.6103    704.4956   636.9924
## 39      443.6115    344.6120    542.6111   514.2173
## 40      673.3878    576.9957    769.7800   647.1938
## 41      570.2727    474.5358    666.0095   542.2178
## 42      580.8009    485.2445    676.3574   602.7555
## 43      416.4709    317.1830    515.7588   426.6896
## 44      595.1849    497.8485    692.5214   619.4148
## 45      617.4601    520.4925    714.4276   614.6528
## 46      541.8615    444.8110    638.9119   613.9141
## 47      676.4960    579.2430    773.7489   625.6636
## 48      579.3539    482.2246    676.4832   573.5844
## 49      569.4618    472.7374    666.1862   586.5229
## 50      709.7126    611.6384    807.7868   756.8776
## 51      595.1391    498.5186    691.7596   624.1171
## 52      592.9066    497.2736    688.5396   582.9542
## 53      503.3138    407.2916    599.3359   436.2796
## 54      556.8992    459.5225    654.2760   607.9046
## 55      630.1715    531.8381    728.5048   712.2499
## 56      707.1764    609.7955    804.5573   685.6361
## 57      464.5831    366.5595    562.6067   380.3482
## 58      480.4624    383.2769    577.6478   430.3903
## 59      753.6936    655.9392    851.4480   736.5829
## 60      593.3173    497.1295    689.5050   541.9331
## 61      456.3902    360.6322    552.1481   517.4296
## 62      540.6680    442.0177    639.3184   498.8789
## 63      520.3144    424.7265    615.9024   564.8867
## 64      655.4442    559.2481    751.6403   651.6743
## 65      579.7023    483.7132    675.6913   537.0380
## 66      564.3485    468.2970    660.3999   537.9019
## 67      626.3723    529.5365    723.2081   617.7234
## 68      647.5179    551.7839    743.2519   674.5304
## 69      638.0141    540.3637    735.6645   595.2126
## 70      458.2966    361.6706    554.9225   413.8889
## 71      605.4607    509.2270    701.6943   549.5685
## 72      599.1682    500.7582    697.5783   658.3994
## 73      520.0244    424.2064    615.8424   531.0190
## 74      494.3730    398.4272    590.3188   460.4072
```

```
## 75      616.5445    519.1282    713.9607   702.5505
## 76      635.7650    539.0089    732.5210   664.8437
## 77      633.2539    537.7824    728.7254   639.9483
## 78      691.9819    594.6167    789.3471   708.1774
## 79      587.2836    487.9345    686.6327   624.8101
## 80      626.3687    529.8652    722.8723   598.1557
## 81      668.3728    571.4569    765.2887   679.2970
## 82      498.9740    398.9434    599.0046   442.9465
## 83      601.4376    505.8594    697.0158   597.2687
## 84      535.3456    434.7090    635.9822   649.0334
## 85      629.4475    532.3392    726.5558   548.9310
## 86      675.6211    579.5005    771.7417   610.8058
## 87      604.5626    507.3408    701.7844   581.8715
## 88      620.0625    523.9965    716.1284   628.3831
## 89      588.0113    490.1754    685.8473   631.9839
## 90      686.5627    583.3667    789.7588   703.6629
## 91      606.1538    508.0804    704.2273   563.2154
## 92      531.5777    434.7543    628.4011   532.6827
## 93      632.2693    535.1817    729.3570   658.7866
## 94      661.7449    565.2655    758.2244   676.5573
## 95      714.0289    617.4201    810.6377   737.1775
## 96      549.5583    452.7399    646.3767   579.5870
## 97      613.5816    515.0295    712.1337   623.8587
## 98      548.2882    451.9282    644.6481   491.1756
## 99      620.2451    522.1987    718.2916   578.8728
## 100     710.7854    612.5917    808.9791   787.7449
## 101     511.8170    416.2245    607.4096   544.3827
## 102     601.0114    503.7648    698.2581   638.1359
## 103     516.1461    420.3210    611.9712   491.9649
## 104     727.9118    632.0715    823.7521   671.4428
## 105     568.7309    470.4473    667.0145   553.3379
## 106     618.1651    520.0165    716.3137   619.2276
## 107     554.3621    456.0958    652.6284   529.6569
## 108     527.3830    430.1321    624.6340   500.8778
## 109     542.8536    445.4874    640.2198   617.2553
## 110     561.7237    465.7295    657.7179   609.5791
## 111     469.7762    373.9268    565.6256   453.4090
## 112     516.8095    420.5075    613.1115   506.9408
## 113     584.4868    486.7974    682.1762   664.9839
## 114     607.1942    510.7300    703.6584   595.6654
## 115     483.8045    387.3163    580.2928   565.0835
## 116     591.7949    494.7232    688.8666   610.0330
## 117     510.8421    413.9152    607.7690   520.4210
## 118     503.6472    404.8051    602.4893   503.6356
## 119     595.6850    499.6276    691.7425   546.5841
## 120     498.3436    400.0958    596.5914   446.8143
## 121     527.8930    430.3598    625.4262   477.8550
## 122     651.2346    554.4802    747.9889   794.6505
## 123     667.5199    569.7875    765.2524   574.2764
## 124     506.8886    411.1048    602.6723   514.5122
## 125     601.8366    505.5179    698.1553   622.8158
## 126     582.7245    485.6211    679.8279   518.5721
## 127     523.7989    427.2724    620.3255   492.2498
## 128     542.0710    442.8251    641.3169   587.7238
```

```
## 129     600.2313    503.8390    696.6236   624.9519
## 130     624.1752    526.7654    721.5851   698.7750
## 131     484.9609    387.9287    581.9931   462.5540
## 132     620.3361    524.8376    715.8345   577.0902
## 133     454.2826    356.8211    551.7442   569.5065
## 134     436.4759    337.8355    535.1162   434.0912
## 135     734.7323    637.9998    831.4649   716.4410
## 136     502.1141    402.3762    601.8520   494.6875
## 137     718.3520    620.4925    816.2115   700.6472
## 138     707.8334    610.3692    805.2976   726.4957
## 139     432.2765    329.6693    534.8837   458.5903
## 140     665.4160    569.0600    761.7720   639.1471
## 141     466.9063    367.8732    565.9395   547.9424
## 142     540.6983    444.1466    637.2500   551.7520
## 143     492.6125    396.8063    588.4187   522.1601
## 144     568.7484    471.8633    665.6335   549.1214
## 145     698.9073    601.4560    796.3586   704.8164
## 146     665.8576    569.4677    762.2475   646.1397
## 147     536.8817    441.1911    632.5723   503.6918
## 148     567.4952    470.9825    664.0080   555.4639
## 149     454.0582    356.6670    551.4495   433.0654
## 150     535.2990    438.2442    632.3539   507.6927
## 151     616.1485    519.1033    713.1937   548.0493
## 152     698.5211    601.7043    795.3380   736.3741
## 153     626.0148    520.9897    731.0400   521.3435
## 154     618.7004    521.0646    716.3361   589.1671
## 155     703.7080    607.3421    800.0739   749.2599
## 156     449.2545    351.6147    546.8943   584.7224
## 157     609.2119    512.3566    706.0671   564.8569
## 158     550.1804    454.0557    646.3052   569.0468
## 159     616.3902    518.8602    713.9201   629.8096
## 160     566.5545    470.9402    662.1689   513.8385
## 161     508.9498    412.6860    605.2135   527.1536
## 162     455.2180    359.3009    551.1351   482.2035
## 163     419.2786    321.7618    516.7953   436.4769
## 164     636.8767    539.0587    734.6948   646.6033
## 165     554.8222    457.5100    652.1344   622.5434
## 166     547.6992    448.3446    647.0537   596.5075
## 167     608.6173    511.6731    705.5616   665.0947
## 168     504.1788    408.5231    599.8345   491.6161
## 169     607.7702    510.9684    704.5720   583.2814
## 170     511.0990    415.3239    606.8741   606.5028
## 171     595.6453    498.0059    693.2847   584.8640
## 172     622.3756    522.2213    722.5299   635.3163
## 173     542.2080    445.2812    639.1349   522.5450
## 174     644.3270    548.3189    740.3352   678.3172
## 175     544.5838    447.3856    641.7819   457.4744
## 176     717.6419    621.8523    813.4315   695.6921
## 177     784.9967    686.4029    883.5905   764.0372
## 178     534.3521    437.3839    631.3203   413.4732
## 179     554.9386    458.6314    651.2457   624.3904
## 180     669.6159    573.7169    765.5148   641.7423
## 181     551.6739    455.7206    647.6271   512.4366
## 182     725.0205    628.4656    821.5755   707.5731
```

```
## 183     433.3519    336.0903    530.6135   440.1676
## 184     738.3554    642.3501    834.3607   739.5392
## 185     655.9110    558.3499    753.4721   694.3444
## 186     679.9263    584.1010    775.7515   763.0419
## 187     545.2586    447.9477    642.5696   548.6716
## 188     630.9500    535.1712    726.7287   704.5717
## 189     581.8916    485.0386    678.7447   551.4120
## 190     567.1105    468.3097    665.9114   517.9553
## 191     578.6743    481.9114    675.4372   638.0033
## 192     562.8784    466.8617    658.8950   533.7296
## 193     594.3014    497.7050    690.8978   655.0662
## 194     524.6465    428.7922    620.5007   484.9591
## 195     465.0755    366.1594    563.9916   475.7100
## 196     661.7267    564.9618    758.4917   727.0107
## 197     534.8946    438.5219    631.2673   428.9199
## 198     565.3133    467.9877    662.6389   572.6965
## 199     639.8759    544.0866    735.6652   761.7539
## 200     505.7321    409.0360    602.4283   480.8677
```

```
## 'data.frame':    191 obs. of  4 variables:
##  $ fitted_value: num  695 589 643 584 620 ...
##  $ lower_bound : num  597 493 548 487 523 ...
##  $ upper_bound : num  793 685 739 682 717 ...
##  $ true_value  : num  659 585 640 605 620 ...
```

Interpretation:

After filtering, we see that 191 of 200 ($191/200 = 95.5\%$) values of CreditScore are in the 95% predictive intervals given by our model *fit5*. We can say that 95.5% individuals lie in the predictive interval, which is a high rate. However, we could not infer causality here.

```
## [1] "num of individual correctly classified = 174.000000"
```

The proportion of individuals who are correctly classified by our model *fit5* is $174/200 = 87\%$.

All in all, our model *fit5* is sufficient.

Summary

The model we build is: $CreditScore \sim Status + (log(Duration))^2 + History + Purpose + (log(Amount))^2 + Savings + Disposable + Personal + OtherParties + log(Age) + Plans + Housing + Telephone + Foreign$ and corresponding coefficients are:

```
##           (Intercept)        StatusNegative            StatusNone
##          642.67278580          -50.33864585           37.80254324
##           StatusSmall         I(Duration^2)               HistoryB
##          -30.63848211          -5.05928694          -1.71064194
##              HistoryC              HistoryD              HistoryE
##           47.43690750           48.61099645           80.96959777
##       PurposeDomestic     PurposeEducation     PurposeFurniture
##          -24.04206997          -53.05132330          -0.65449404
##         PurposeNewCar         PurposeOther       PurposeRepairs
##          -46.39521729          -17.52376866         -26.72275168
##     PurposeTelevision      PurposeTraining       PurposeUsedCar
##            2.33951122           60.86332382           37.63425410
##           I(Amount^2)            SavingsLow        SavingsMedium
##           -1.50658221          -13.21931180            0.09274299
```

```
##        SavingsUnknown      SavingsVeryLarge           Disposable2
##           25.96072737           61.98825705          -12.53188250
##           Disposable3           Disposable4      PersonalF:Single
##          -27.68161822          -42.74550056             7.66456101
##   PersonalM:DivSepMar      PersonalM:Single OtherPartiesGuarantor
##          -13.69298484           29.91992387           61.43832784
##      OtherPartiesNone                   Age             PlansNone
##           12.31840746           16.45256118           32.60000098
##           PlansStores           HousingRent       HousingRentFree
##           12.05621218          -17.08723904           -2.90474034
##          TelephoneYes             ForeignYes
##           12.83855854          -54.70072304
```

We found strong evidence between CreditScore and factors like Status, History, Purpose, Savings, Personal, Disposable, OtherParties, Plans, Housing, Telephone, Foreign. Specifically, applicant with Status (None) or History (E) or Purpose (Training) or Savings (Verylarge) or Disposable (3) or Personal (Single) or OtherParties (Guarantor) or Plans (None) or Housing (Own) or Telephone (Yes) or Foreign (No) has a higher CreditScore, which could be the basis of lending decision. Moreover, we see that applicant who requests lower amount with shorter duration seems to have higher credit score and applicant who requests lower amount with identity of a non-foreign worker has a higher credit score and applicant with worse previous loan history has higher credit score, which might due to correlation between History and Amount, but the causality is to be determined. For continuous variables, Age has a strong negative association with CreditScore while Duration and Amount have strong negative association with CreditScore, but the causality is to be determined.

After prediction, we found our model is much better than the full model by considering the MSE. We also found our 95% predictive intervals contains 191 among 200 samples, which suggests a considerable match. The proportion of the individuals correctly classified is 87%.

In conclusion, Duration, Amount and Age therefore seem the most plausible explanation for differences in CreditScore. We note the significant association between the other variables and CreditScore when considered on their own. Moreover, our model could be improved by considering the correlations between factors and continuous variables, such as Foreign and Duration, History and Amount, Duration and Amount, etc. For instance, adding interactions between variables might increase the accuracy of our model.

Finally, there could be other important variables affecting credit score that we haven't observed and a causal link can not be inferred based on this study alone, more samples and further tests are required.